Forschungsbereich Genetik und Biometrie
Leibniz-Institut für Nutztierbiologie, FBN Dummerstorf

A systems biological approach towards the molecular basis of heterosis in *Arabidopsis thaliana*

Dissertation

zur Erlangung des akademischen Grades
"doctor rerum naturalium" (Dr. rer. nat.)

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät
der Universität Potsdam

von Sandra Andorf

Potsdam, den 09.11.2010

Published online at the
Institutional Repository of the University of Potsdam:
URL http://opus.kobv.de/ubp/volltexte/2011/5117/
URN urn:nbn:de:kobv:517-opus-51173
http://nbn-resolving.org/urn:nbn:de:kobv:517-opus-51173

Contents

Abstract

Heterosis is defined as the superiority in performance of heterozygous genotypes compared to their corresponding genetically different homozygous parents. This phenomenon is already known since the beginning of the last century and it has been widely used in plant breeding, but the underlying genetic and molecular mechanisms are not well understood.

In this work, a systems biological approach based on molecular network structures is proposed to contribute to the understanding of heterosis.

Hybrids are likely to contain additional regulatory possibilities compared to their homozygous parents and, therefore, they may be able to correctly respond to a higher number of environmental challenges, which leads to a higher adaptability and, thus, the heterosis phenomenon.

In the network hypothesis for heterosis, presented in this work, more regulatory interactions are expected in the molecular networks of the hybrids compared to the homozygous parents. Partial correlations were used to assess this difference in the global interaction structure of regulatory networks between the hybrids and the homozygous genotypes.

This network hypothesis for heterosis was tested on metabolite profiles as well as gene expression data of the two parental *Arabidopsis thaliana* accessions C24 and Col-0 and their reciprocal crosses. These plants are known to show a heterosis effect in their biomass phenotype. The hypothesis was confirmed for mid-parent and best-parent heterosis for either hybrid of our experimental metabolite as well as gene expression data. It was shown that this result is influenced by the used cutoffs during the analyses. Too strict filtering resulted in sets of metabolites and genes for which the network hypothesis for heterosis does not hold true for either hybrid regarding mid-parent as well as best-parent heterosis.

In an over-representation analysis, the genes that show the largest heterosis effects according to our network hypothesis were compared to genes of heterotic quantitative trait loci (QTL) regions. Separately for either hybrid regarding mid-parent as well as best-parent heterosis, a significantly larger overlap between the resulting gene lists of the two different approaches towards biomass heterosis was detected than expected by chance. This suggests that each heterotic QTL region contains many genes influencing biomass heterosis in the early development of *Arabidopsis thaliana*. Furthermore, this integrative analysis led to a confinement and an increased confidence in the group of candidate genes for biomass heterosis in *Arabidopsis thaliana* identified by both approaches.

Abbreviations

BN Bayesian network

BPH best-parent heterosis

CIM composite interval mapping

cM centimorgan

DAS days after sowing

FDR false discovery rate

GC gas chromatography

GGM graphical Gaussian model

GO Gene Ontology

GSEA Gene Set Enrichment Analysis

HAS hours after sowing

KEGG Kyoto Encyclopedia of Genes and Genomes

MPH mid-parent heterosis

MS mass spectrometry

ORA over-representation analysis

PO Plant Ontology

QTL quantitative trait loci

RIL recombinant inbred line

RN relevance network

SAGE serial analysis of gene expression

SNP single nucleotide polymorphism

TAIR The Arabidopsis Information Resource

1 Introduction

In this work, a systems biological approach to gain a little further insight into the molecular basis of biomass heterosis in *Arabidopsis thaliana* plants is presented. For this basic research a "network hypothesis for heterosis" is proposed and tested on experimental data from different omics levels.
Furthermore, the systems biological analysis is integrated with results from a quantitative genetics approach towards biomass heterosis in *Arabidopsis thaliana*. It is tested if two different approaches point to similar genomic regions influencing this trait under study.
In this introduction, first, the heterosis phenomenon is explained. Afterwards, basic concepts, methods and the experimental data are introduced briefly. The network hypothesis for heterosis builds the basis for this work and is presented in section 1.7.

1.1 Heterosis

Heterosis, also known as hybrid vigor, is defined as the superior performance of heterozygous genotypes compared to their homozygous parental inbred lines (Shull, 1952). The superiority of the hybrids is expressed as increased biomass, size, yield, speed of development, fertility, resistance to disease or to insect pest (Birchler et al., 2003; Hochholdinger and Hoecker, 2007). This phenomenon can be detected in animals as well as in plants.
Heterosis is either measured as mid-parent heterosis (MPH), defined as the difference between the hybrid and the mean of the parents, or as best-parent heterosis (BPH), the deviation of the trait value of a hybrid from the better parent (Falconer and Mackay, 1996; Lamkey and Edwards, 1999).
The molecular basis of heterosis is still unknown but three different models were proposed to explain the phenomenon. The two classical quantitative genetic explanations include the *dominance* and the *overdominance* model (Crow, 1948). The dominance hypothesis explains heterosis as the complementation of deleterious alleles by favorable dominant alleles from the other parent at multiple loci in the hybrid (Davenport, 1908; Bruce, 1910). The overdominance hypothesis states that

interactions of different alleles occur at one or multiple heterozygous loci that lead to hybrids that perform better than either homozygous parent (Hull, 1945; Crow, 1948).
A variation of the dominance hypothesis is the so-called *pseudo-overdominance* model. This model explains the situation that recessive alleles of tightly linked genes, which are located on opposite homologs, are complemented by the superior dominant alleles in the hybrid. In this situation it seems like overdominance is operating, which resulted in the name pseudo-overdominance (Crow, 1952; Birchler et al., 2010).
Finally, the epistasis hypothesis explains heterosis by interactions of favorable genes located at two or more different loci (Powers, 1944; Williams, 1959).
The interest to understand the basis of heterosis is high because its use plays an important role in plant and animal breeding to maximize the agronomic performance (Melchinger et al., 2007b). In 2007, around 95% of the U.S. corn acreage and 65% of the corn acreage worldwide were planted to hybrids (Hochholdinger and Hoecker, 2007).
Even though the heterosis phenomenon is known for more than 100 years and it is widely used in plant breeding, the underlying genetic and molecular mechanisms are not yet established.
Different approaches following the target to gain deeper insight into the molecular mechanisms of heterosis exist. Quantitative genetics approaches (e.g. Frascaroli et al., 2007; Melchinger et al., 2007a; Meyer et al., 2010) as well as analyses of functional data such as gene expression profiles of selected genes or pathways (e.g. Meyer et al., 2007; Thiemann et al., 2010) or all genes of an organism (e.g. Swanson-Wagner et al., 2006; Frisch et al., 2010) were performed on different species and developmental stages to evaluate various traits. Other approaches aim to *predict* heterosis in the hybrids using genetic and/or functional characteristics of the parental lines (e.g. Gärtner et al., 2009; Frisch et al., 2010; Schrag et al., 2010). Different studies led to different conclusions, suggesting dominance, overdominance (or pseudo-overdominance) as well as epistasis and all possible combinations out of these as the predominant mechanism underlying heterosis.

1.2 Systems biology

To understand the complex events within or between biological cells, it is not enough to identify and study the single components that the cells are built of. It is rather important to analyze these components together as a system. The structure as well as the dynamics of all parts in the system have to be examined to be able to explain the

functioning of the whole biological system (Kitano, 2002). This systematic, holistic view of biological organisms is the focus of *systems biology*. Nevertheless, up to now no concise definition of systems biology exists. It can be viewed as the integration of genomics, proteomics and other omics data to understand the interplay of different components and get insight into biological processes on the systems-level (Hood and Galas, 2003; Ge et al., 2003; Westerhoff and Palsson, 2004). The large sizes of the datasets, the nonlinear character of the interactions that have to be elucidated and the resulting complexity of the cellular system make the use of mathematical models from systems theory essential to describe the structure and dynamic of a biological system (Wolkenhauer, 2001; Ideker et al., 2001; Wolkenhauer, 2007). So, systems biology can be seen as the simulation and study of complex biological processes by integrating genome-wide experimental data and mathematical modeling approaches. Since there is no exact definition of systems biology, I want to point out that in my opinion the systems biological character of the work presented here is given by the integration of two different omics levels (metabolomics and transcriptomics; explained in the next section). Not single genes or metabolites are studied and compared for different genotypes but the estimated underlying global regulatory networks. Therefore, the focus in this work is on differences on a systems level and not in single components.

1.3 Omics techniques

In molecular or cell biology only a few features (e.g. genes, proteins or metabolites) are studied at a time. However, as already pointed out in the systems biology section, genes (and their products) do not function alone but in combination with each other and, therefore, they have to be analyzed as a whole to get more insight into biological processes (Hartwell et al., 1999; Ge et al., 2003). The development of high-throughput techniques, such as DNA and protein microarrays or mass spectrometry, made it possible to measure the *whole of the objects* of one level of gene expression simultaneously. This *whole of objects* are all gene products of one level of gene expression that are present in a biological sample. It is labeled with the suffix "-ome", while the terms that name the field of biology that aims to study these "omes" end on "-omics". Omics technologies follow the goal to identify all gene products of one level in a biological sample along with their properties and quantitative dynamics (Weckwerth, 2003).

The *genome* is the entire set of DNA sequence information of an organism and the belonging field of study is called *genomics*. Along with the "popularity" of genomics,

a variety of omics subdisciplines has begun to emerge. One example is proteomics, the study of all proteins of a cell with their particular modifications at a given time point. Unlike the genome, which is the same in all cells of one organism at each time, the proteome is different in every cell type and changes over the time and with varying conditions (Campbell and Heyer, 2002).
Because data of the two omics levels of metabolomics and transcriptomics are analyzed in this work, they are explained in more detail in the next sections.

1.3.1 Metabolomics

The metabolome is the quantitative complement of all metabolites expressed in a biological sample under particular conditions (Oliver et al., 1998; Kell et al., 2005). Metabolomics aims to identify and quantify the metabolome. The main experimental technique used to achieve this is the coupling between gas or liquid chromatography with mass spectrometry. These techniques allow for the analysis of dynamic systems and make metabolomics a key technology in systems biology (Weckwerth, 2003).
However, no single metabolomics technique is able to measure *all* low-molecular-weight metabolites. In reality, datasets comprise only a small probably biased fraction of all metabolites.
As shown by Fiehn et al. (2000), metabolome analyses can be used to enhance the power of existing functional genomic approaches to describe functions and interactions of genes or proteins. Furthermore, metabolomic algorithms make it possible to model and reconstruct small metabolic networks that show the relations between metabolites. The study of metabolic networks is a major step to the understanding of complex biochemical systems and living organisms and plays an important role in the discovery of drug targets (Kell, 2004; Guimerà and Amaral, 2005).

1.3.2 Transcriptomics

The transcriptome is the collection and quantity of all transcripts in one cell or population of cells at any given time point (Campbell and Heyer, 2002; Wang et al., 2009). The transcription of genes is the first step in gene regulation. It varies in different cell types and changes with the stage of development or environmental conditions. So, such as the proteome, the transcriptome is extremely dynamic (Velculescu et al., 1997). Its study, the transcriptomics, is important to understand genes and pathways involved in biological processes. The analysis of gene expression patterns follows multiple aims of which only two are presented here. Transcriptomics can be used to study the changing expression levels of each transcript under different conditions

(Wang et al., 2009). This can be the basis for the identification of possible drug targets or diagnostic biomarkers, for example, by comparing which genes are highly expressed in tumorous but not in healthy tissues. Furthermore, transcriptomics can be helpful to get a further insight into the functional annotation of genes based on the assumption that genes belonging to the same regulatory pathway are more highly co-expressed than genes from different pathways (Wei et al., 2006). So, genes with similar expression patterns are likely to be functionally related and controlled by the same molecular regulatory mechanism.

The most common techniques for the analysis of gene expression data are different kinds of microarrays, which are hybridization-based techniques, and sequence-based approaches like SAGE (serial analysis of gene expression) or novel high-throughput DNA sequencing methods termed RNA-Seq (Wang et al., 2009).

1.4 Quantitative trait loci

In general, it is differentiated between discrete and quantitative traits. Discrete traits are present in several distinct characteristics, e.g. colors of flowers. Quantitative traits, such as yield, are measurable on a continuous scale. Geneticists use quantitative traits (specific phenotypes) to infer the underlying genetics.

Quantitative trait loci (QTL) are segments of a chromosome, which have an influence on the quantitative phenotypic trait of an organism. Polymorphic genetic markers (markers which show at least two different alleles at a locus) are used for QTL mapping. Frequently used markers are single nucleotide polymorphism (SNP) or mini-/microsatellites.

The coupling between marker and locus which is important for the trait under study builds the basis for the identification of QTL. The underlying genetic principle is that loci that are located at physically close chromosomal regions (linked loci) show a higher probability to be passed together from one generation to the next one than distant loci. This is based on the fact that the probability of a recombination increases with increasing distance between loci (Campbell and Heyer, 2002; Mount, 2004).

QTL mapping experiments are important in human genetics, e.g. for the identification of genes that cause diseases, but also for plant and animal breeding (e.g. Hackett, 2002).

1.5 Reverse engineering

Reverse engineering is one approach in the field of systems biology. Depending on the definition, it can also be placed between systems biology and bioinformatics (Hache et al., 2009). The general goal of reverse engineering methods is to identify a model of the inner workings of a system by analyzing the observable outputs which are based on the interplay of the single objects in the system (Ingolia and Weissman, 2008). These approaches are not only applied in biology but in many different areas, such as mechanical or software engineering. In molecular biology, reverse engineering methods aim to use experimental data to reconstruct the structure of the underlying unknown biological network. If this is possible, the inferred networks can be studied to increase the understanding of cellular functions (Brazhnik et al., 2002).

The opposite of reverse engineering algorithms build the so-called forward modeling approaches, which try to predict gene expression profiles on the basis of known gene regulatory networks and their dynamics.

The biological networks which structures are elucidated in a reverse engineering approach can be very different. In general, networks consist of nodes, representing some kind of objects, and edges that connect the nodes and symbolize the relation between the adjacent objects (nodes). In gene regulatory networks the nodes are genes while the edges are transcriptional regulatory interactions. However, in other studies an edge between two genes can also symbolize that the genes are co-regulated, participate in a common pathway, share a common biological function or in the case of a directed edge it may represent a step in a metabolic pathway, signal transduction cascade or stage of development. So, the edges of reconstructed biological networks have to be interpreted with care and with respect to the applied mathematical model (Hartemink, 2005; Ma et al., 2007).

The reliability of the predicted networks and which approach should be used for a particular scientific problem are still challenging questions (Gardner and Faith, 2005). Due to their availability, especially gene expression profiles are used as input for reverse engineering approaches to reconstruct gene regulatory networks (Hache et al., 2009).

The basic principle of reverse engineering algorithms to infer the underlying biological networks or biochemical pathways from genome-wide experimental data is the same for each approach. First, the experimental data is obtained. The same variables (features such as genes or metabolites) have to be observed several times (as a time series; under different conditions or treatments etc.). This profile data can be observational or interventional. In modern molecular biology there exist many

different techniques of perturbations or interventions, such as knock out experiments (Werhli et al., 2006). Afterwards, a mathematical algorithm is applied to determine a model that describes the regulatory system underlying the observed data (Gardner and Faith, 2005).
Several different reverse engineering approaches were proposed that vary in the mathematical model and the algorithm that is used for the inference of biological networks. These mathematical models include but are not limited to linear models (D'haeseleer et al., 1999), differential equations (de Jong, 2002), static or dynamic Bayesian networks (BNs) (Friedman et al., 2000; Imoto et al., 2003), relevance networks (RNs) (Butte et al., 2000; Basso et al., 2005) and association networks (often described by graphical Gaussian models (GGMs)) (Kishino and Waddell, 2000; Opgen-Rhein and Strimmer, 2007b).
GGMs build one basis of the "network hypothesis for heterosis" which is presented in section 1.7. Therefore, in the following the main focus will be on GGMs and only a briefly comparison to RNs and BNs is presented.

RNs are currently one of the most widely used mathematical models to infer underlying biological networks from high-throughput omics data. They are based on pairwise association scores between all investigated features. Pearson correlations or mutual information were proposed as good association scores (Butte et al., 2000; Basso et al., 2005). The construction of biological networks out of e.g. gene expression data is straight forward using standard Pearson correlations. If the association score for the profiles of a pair of genes exceeds a preselected threshold, a functional interaction, influence or dependency is assumed between these two genes (Ma et al., 2007).
While this approach is easy and computationally not expensive, it has the disadvantage that the interactions between two features are calculated without involving the other features of the system. So, correlation networks can not distinguish between direct and indirect interactions. An indirect interaction might emerge between two features if these features are uncorrelated among themselves but highly correlated to a common third feature. Consequently, nearly all features will be correlated and, therefore, connected to each other in the resulting network. Only a missing edge (zero correlation) provides information (indicating independence) but not the presence of an edge. However, the dependence between feature is what is important to understand complex biological functions and not the independence that can be identified using RNs (Schäfer and Strimmer, 2005b). Furthermore, in the case of heavily connected networks, reverse engineering algorithms based on RNs would

probably lead to ambiguous results. For these reasons, RNs are of limited use to study biological networks (Brazhnik et al., 2002; Schäfer and Strimmer, 2005a; Werhli et al., 2006).

An alternative to RNs are GGMs. GGMs are undirected probabilistic graphical models. These models allow to differentiate between direct and indirect interactions (Schäfer and Strimmer, 2005a; Opgen-Rhein and Strimmer, 2007b). Therefore, using GGMs makes the reduction of the number of indirect interactions in the inferred network possible. Hence, GGMs can be used to study *dependencies* between features but offer only a weak criterion of *independence*.

Kishino and Waddell (2000) were the first who proposed GGMs to model association structures between genes. In GGNs, partial correlations are used to evaluate direct interactions between all pairs of features under study (Toh and Horimoto, 2002). A partial correlation between two variables is the correlation that remains between the two variables after the effect of all other variables has been subtracted. This removal of the effects of the other variables can be done on the basis of a linear regression of each of the two variables to all remaining variables (Opgen-Rhein and Strimmer, 2007b).

From standard graphical model theory it is known that the partial correlation matrix is related to the inverse of the standard covariance matrix. This statement is equally valid for the inverse of the correlation matrix (Schäfer and Strimmer, 2005a). However, the calculation of the partial correlation matrix of all features based on the inverse of the covariance matrix is only possible if the number of observations (samples) is larger than the number of variables (features) (Kishino and Waddell, 2000). If this is not the case, partial correlations cannot be calculated because of not positive definite sample covariance and correlation matrices (Friedman, 1989). Unfortunately, large-scale data, such as gene expression data, usually consist of a large number of variables and a much smaller number of observations, because e.g. on each microarray thousands of genes can be measured but the number of microarrays used is limited due to the prize and manpower. Over the last years several reverse engineering algorithms for many variables and few observations based on GGMs were proposed (e.g. Magwene and Kim, 2004; Wille et al., 2004; Schäfer and Strimmer, 2005b).

The approach by Wille et al. (2004) is to apply GGMs not to the complete set of features but to estimate the dependence between two features conditional on only one other gene. The resulting subnetworks are combined to the complete network. Such an algorithm using limited order partial correlations was also proposed by

Magwene and Kim (2004) and de la Fuente et al. (2004). These limited order partial correlation approaches, however, lead to network models more similar to RNs than to association networks. Hence, the problems of RNs also apply partly for these algorithms.

Other approaches to use GGMs to infer the network structure from small sample data are based on the introduction of regularization and moderation. The goal is to develop estimation methods for the covariance matrix and its inverse that are applicable for a small samples size and many features (Ma et al., 2007).
Such an approach was proposed by Schäfer and Strimmer (2005b). It is based on a shrinkage estimation of the partial correlations. Applying this shrinkage approach, it is possible to estimate the inverse of the covariance matrix as well as correlation matrix for small sample sizes but many features. The basic principle is to shrink the unrestricted sample correlation matrix towards some target. The most commonly used shrinkage targets are the identity matrix or its scalar multiple. The shrinkage estimate is a linear combination (weighted average) of the empirical correlation matrix of the sample and the target. The weight in this weighted average is referred to as the shrinkage parameter. A shrinkage parameter of 1 leads to a shrinkage estimate that equals the shrinkage target. Correspondingly, if the shrinkage parameter equals 0, no shrinkage occurs and the sample correlation matrix is recovered. This shrinkage parameter can be set to one fixed value. However, it is better to select the optimal value for the shrinkage parameter by minimizing a risk function such as the mean squared error (Schäfer and Strimmer, 2005b; Opgen-Rhein and Strimmer, 2007a).
Several approaches were proposed to estimate the minimizing shrinkage parameter, for instance cross-validation (e.g. Friedman, 1989) or empirical Bayes approaches (e.g. Greenland, 2000). As shown by Ledoit and Wolf (2003) and Schäfer and Strimmer (2005b), an analytical determination of this parameter to minimize the mean square error is also possible.
The next step in inferring biological networks, after estimating the partial correlations using a shrinkage approach, is the identification of statistically significant edges in the GGM network. In the approach by Schäfer and Strimmer (2005b), this is done by fitting a mixed linear model to the estimated partial correlations. This way, two-sided P-values corresponding to the null hypothesis of zero partial correlation can be computed (Schäfer and Strimmer, 2005a; Schäfer et al., 2006; Strimmer, 2008).
This shrinkage approach to GGMs is implemented in the *R* package *GeneNet* (Schäfer et al., 2006; Opgen-Rhein and Strimmer, 2007b). Schäfer et al. (2006) chose in

their approach to shrink the empirical correlations towards the identity matrix. The shrinkage parameter is estimated using an analytic formula according to Schäfer and Strimmer (2005b), leading to a distribution-free shrinkage estimation (Opgen-Rhein and Strimmer, 2007a).

Werhli et al. (2006) compared the two reverse engineering methods described above to BNs. BNs describe causal interactions using directed acyclic graphs. Each BN is defined by a graphical structure (the topology) and a family of (conditional) probability distributions (Husmeier, 2003). The nodes represent random variables and the edges conditional dependence relations (Hache et al., 2009). To identify the network structure that is most supported by the experimental data, the mode of the posterior probability has to be determined (Husmeier, 2003). However, nearly all learning algorithms based on directed acyclic graphs show the same problem as GGMs that they were developed for comparatively small numbers of variables and large sample sizes (Tsamardinos et al., 2006). For experimental data with many features and only a few measurements, no single network structure can be adequately identified that represents the posterior probability mode (Husmeier, 2003). In recent years, several algorithms have been proposed that overcome this problem for using BNs for learning network structures from high-dimensional experimental data (e.g. Imoto et al., 2003; Friedman, 2004; Beal et al., 2005).

Reverse engineering algorithms based on BNs are computationally very expensive compared to GGMs and RNs. Werhli et al. (2006) have shown that GGMs as well as BNs outperform RNs with regard to the accuracy of reconstructing gene regulatory networks from high-throughput data. However, no significant difference between GGMs and BNs was found for observational data. Werhli et al. (2006) used in their comparison the GGM approach based on Schäfer and Strimmer (2005b). It led to similar accuracy as computationally much more demanding methods based on (dynamical) BNs. Therefore, the *R* implementation of the reverse engineering algorithms based on GGMs by Opgen-Rhein and Strimmer (2007b) is shown to be appropriate to infer gene regulatory networks from observational omics data.
While Werhli et al. (2006) have shown that the use of GGMs is a good tool to infer gene regulatory networks from observational *gene expression data*, Çakır et al. (2009) concluded in a similar comparative study that partial correlations are momentarily also the best available approach for the inference of metabolic networks from observational *metabolic data* at steady state.

1.6 Enrichment analyses

Many studies aim to optimize analytical techniques to accurately identify a biologically interesting group of genes (e.g. differentially expressed genes in microarray experiments) and to determine their statistical significance. Nowadays, the difficulty is not to determine these genes but their biological function.

Furthermore, the comparison of results from different experiments, under different conditions and at various layers of regulation is difficult. One approach to overcome these problems is to apply an enrichment analysis. Enrichment analyses in general compute whether two lists of features show a statistically significant overlap (Ackermann and Strimmer, 2009; Lachmann and Ma'ayan, 2010).

In enrichment analyses a list of experimentally identified "interesting" genes (or any other features) are analyzed regarding their membership in a-priori defined gene sets. In most applications, gene sets group all genes belonging to one functional annotation or chromosomal location together (Goeman and Bühlmann, 2007). The annotation is based on various databases in different studies. Most common is to use the Gene Ontology (GO) database for gene sets with specific functional categories (Lachmann and Ma'ayan, 2010). Other databases used to annotate gene sets are pathway databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000) or GenMAPP (Doniger et al., 2003).

Enrichment analyses make it possible to interpret the results of high-throughput screening experiments regarding the annotation criterion that was used for the gene sets. In case that each gene set is build out of all known genes belonging to one biological pathway, an enrichment analysis can e.g. determine the activity of a certain pathway under different conditions. So, enrichment approaches make use of previous biological knowledge in the interpretation of omics experiments. Because of this and the fact that lists of genes are analyzed instead of single determined genes alone, enrichment analyses can be used to yield biologically more meaningful interpretations of results of omics studies than approaches based on single genes. This, in turn, makes a better understanding of functional mechanisms in cells possible (Goeman and Bühlmann, 2007; Ackermann and Strimmer, 2009).

Over the last years, several different statistical procedures for enrichment analyses have been introduced. A simple and probably most popular method is the over-representation analysis (ORA). In ORA, first a list of interesting genes, e.g. differentially expressed genes, (or any other features) is determined. This can be done in many different ways, e.g. on the basis of fold change, t-statistics, (shrinkage) correlation coefficient or log-likelihood ratio. Additionally, in most cases, these statistics

are transformed (e.g. absolute values, squared values, ranks, P-values). A strict cut-off is applied to these gene-wise measures to determine a list of top ranked genes (Ackermann and Strimmer, 2009).

Afterwards, it is tested if the genes of a gene set are overrepresented in this list considering all genes analyzed in the particular experiment. This is done by testing a 2×2 contingency table: the independence of the membership of genes in a functional category (gene set) and the membership in the list of top ranked genes is tested (Drăghici et al., 2003; Khatri and Draghici, 2005). Several statistical tests were proposed to test the significance of the overlap between pairs of gene lists, such as the binomial proportions, Chi-squared or Fisher's exact test (Rivals et al., 2007; Lachmann and Ma'ayan, 2010). Any two lists of genes may overlap just by chance and these contingency table statistics can be used to determine gene sets that show an unexpected significant overlap (an overlap that deviates from what is expected by chance) between the list of genes of interest and the genes in the gene set (Backes et al., 2007; Lachmann and Ma'ayan, 2010). In the case that significantly more genes are in the overlap than expected just by chance, the genes of the gene set are called enriched (over-represented) in the list of interesting genes. If significantly less genes are observed in the overlap than expected by chance, the genes are under-represented.

In ORA any kind of list of interesting features, no matter how it was determined, can be assessed regarding its biological background. ORA require no ranking of the interesting features. While this can be an advantage compared to some other methods, it is at the same time a problem of this enrichment approach when applied to e.g. gene expression data. Some authors criticized the requirement of a strict cutoff and that no clear way to determine the list of interesting features that is tested against the gene sets exists (Goeman and Bühlmann, 2007).

Alternatives have been proposed that use the whole vector of the gene-wise measures (e.g. the P-values) instead of only a list of interesting features applying one particular cutoff (Goeman and Bühlmann, 2007). One popular alternative, called "Gene Set Enrichment Analysis" (GSEA), was introduced by Mootha et al. (2003) and improved by Subramanian et al. (2005). Different from the ORA, this algorithm takes the ranking of the interesting genes (or other features) according to their correlation to the phenotype of interest (e.g. the t-test statistic) into account. This sorted list of genes is processed from top to bottom determining a running sum for each gene set. This running sum for one gene set is increased every time a gene from the sorted list of interesting genes belongs also to the gene set. Respectively, the running sum is decreased every time the gene is not in the gene set. The magnitude of the increment in each running step depends on the rank and calculated gene-wise measure of the

individual gene. The maximum deviation of the running sum from zero is the so-called enrichment score. This enrichment score can be used to test the null hypothesis that genes in the gene set are randomly spread in the list of all genes under study in the particular experiment (Ackermann and Strimmer, 2009). The significance of an observed enrichment score is computed as the probability that a random running sum reaches a value that is as high as the observed enrichment score (Backes et al., 2007).

1.6.1 Integrative analyses

Enrichment analyses are mainly used to analyze results of high-throughput screening experiments regarding their enrichment in gene sets that are annotated based on biological background knowledge stored in databases. However, these techniques can also be used to integrate the results from different experiments.
Each omics technique addresses only one concrete level of biological organization, e.g. transcripts, metabolites or proteins. Furthermore, every technique adds its own technical variance and methodological bias to its results (Steinfath et al., 2007). These two problems can be reduced by an integrative analysis of two or more distinct omics analyses. The integrative analysis of different levels of biological organization can increase the understanding of the functional activities in and between living cells and help to formulate biological hypotheses (Ge et al., 2003).
If two different omics techniques applied to answer the same biological question, point to similar candidate genes, the functional relevance of this result can be established with increased confidence (Steinfath et al., 2007).
In the case that an enrichment analysis is used to integrate the results from different experiments towards the same biological question, the somehow determined interesting features of the first experiment are used as the gene set. These features are tested for over-representation in the identified interesting features of the second experiment, on the background of all features analyzed in this second experiment.

1.7 Network hypothesis for heterosis

The hypothesis that is presented and tested in this work aims to propose a further understanding of heterosis on the level of molecular network structures. The basis for this hypothesis builds, in general, the understanding of the heterosis phenomenon as increased adaptability. This basic idea was already proposed by Shull (1908) and, in particular, in the work by Robertson and Reeve (1952). In experiments on

Drosophila melanogaster, Robertson and Reeve (1952) discovered that the variance in wing-length, which is highly correlated with body-size, is on average nearly twice as high in inbred lines as it is in crosses between these lines. This findings suggest that the environmental variance of such quantitative traits is smaller in heterozygotes than in homozygous genotypes.

To further analyze the relation between heterozygosity and environmental variance they prepared genotypes that are heterozygous in between zero and three chromosomes. For every degree of heterozygosity the average variance of wing-length of several different genotypes was calculated. The average wing-length variance decreased with increasing heterozygosity. This indicated that the environmental variance of every phenotype is related to the degree of heterozygosity. Robertson and Reeve (1952) detected the same tendency also for size and rate of egg production and suggested that this phenomenon of declining susceptibility to environmental variations for increasing heterozygosity might appear for many quantitative traits in animals and plants. They concluded in their work that individuals with increased heterozygosity carry a greater diversity of alleles which leads to a larger biochemical versatility in development. Based on this greater biochemical versatility, the heterozygous genotypes are able to correctly respond to more environmental variations and make better use of the materials available in the environment than the inbred lines, which, in turn, leads to the heterosis phenomenon.

In the here presented "network hypothesis for heterosis" this findings are extended in the way that it is assumed that more regulatory possibilities in the heterozygous genotypes go along with more regulatory interactions on the molecular level. So, it is hypothesized that the hybrids, which show heterosis, contain denser regulatory networks (more regulatory interactions) than the homozygous parents.

In order to test this assumption, a way of identifying the number of regulatory interactions that are probably present in regulatory networks had to be found. As explained in section 1.5, reverse engineering approaches aim to infer biological networks from omics data. Following Werhli et al. (2006) the approach based on GGMs by Schäfer and Strimmer (2005b), that allows to calculate partial correlations of omics data measured at different time points or for a series of environmental changes or developmental stages, is used in the here presented work to estimate the number of regulatory interactions that are probably present in the regulatory networks of different genotypes.

Summarizing, in the network hypothesis for heterosis, more regulatory interactions are expected in heterozygous genotypes, which show heterosis in their phenotype, than in the regulatory networks of the homozygous parental lines.

In our experimental design of two homozygous parental lines and both reciprocal crosses, partial correlations of the observational profile omics data are used to estimate the probability of additional regulatory interactions in the hybrids.

1.8 Thesis outline

This work takes a systems biological approach based on network structures to gain a little further insight into the basis of heterosis. In particular, biomass heterosis in the early development of *Arabidopsis thaliana* plants is studied. In all analyses the experimental data was observed for the two homozygous lines C24 × C24 and Col-0 × Col-0 of the model plant *Arabidopsis thaliana* and their reciprocal crosses C24 × Col-0 and Col-0 × C24.

The first part of a heterozygous genotype refers to the maternal plant and the second part the genotype of the paternal plant. So, C24 × Col-0 is the cross between C24 × C24 as "mother" plant and Col-0 × Col-0 as the "father". It is important that both reciprocal crosses between two genotypes are studied because in most plants the plastids are only inherited from one parent. In *Arabidopsis thaliana*, the mitochondrial and plastid DNAs are inherited maternally (Martínez et al., 1997; Ruf et al., 2007).

In chapter 2 (published as Andorf et al., 2009), first, a simulation study based on an artificial neuronal network is presented that demonstrates that more regulatory possibilities go along with denser regulatory networks. Furthermore, it is shown in the simulation study that causal interactions lead to increased partial correlations between the two corresponding nodes of the trained artificial neuronal network. These simulation results built a further basis, besides the literature, for the "network hypothesis for heterosis" presented in section 1.7.

Moreover, the test of this hypothesis on metabolite profiles of the before described four *Arabidopsis thaliana* genotypes is presented in chapter 2. Partial correlations according to Schäfer and Strimmer (2005b) are calculated for all four genotypes to estimate the connectivity of the underlying regulatory networks. Based on these partial correlations, "partial correlation mid-parent heterosis" values for each metabolite of either hybrid are calculated. Positive partial correlation heterosis values suggest that the particular metabolite is involved in more regulatory interactions in the hybrid than in the mid-parent expectation. Following the network hypothesis for heterosis, positive partial correlation heterosis values are expected for the majority of the metabolites.

While reverse engineering analyses are very frequently applied in transcriptomics to infer gene regulatory networks, the in chapter 2 presented use of metabolite data to infer biological networks was a relatively untouched area in 2009 (Çakır et al., 2009) when this analysis was published as Andorf et al. (2009).
The network hypothesis for heterosis is also tested on *gene expression data* of the four *Arabidopsis thaliana* genotypes (published as Andorf et al. (2010a)). In this analysis, presented in chapter 3, the significances of the partial correlations build the basis of the calculation of the mid-parent and best-parent heterosis values. This chapter also covers an over-representation analysis to determine if the genes that show the strongest heterosis effects according to the network hypothesis for heterosis are particularly enriched in gene sets that were annotated according to The Arabidopsis Information Resource (TAIR) (Huala et al., 2001) and Plant Ontology (PO) (The Plant Ontology Consortium, 2002) databases.
The metabolite as well as gene expression data was measured at seven time points during the early development of *Arabidopsis thaliana*. These "time series" profiles are not time series from the mathematical or systems biological point of view. The studied time points are several days apart from each other, which is a long part in the life of *Arabidopsis thaliana* plants. In the experiments, data from steady states are analyzed so that the dynamics within the biological system can not be studied.
In the 4th chapter, an integrative analysis of the results of the gene expression data according to the network hypothesis for heterosis and genes that were identified in QTL mapping experiments to influence biomass heterosis in *Arabidopsis thaliana* (Meyer et al., 2010) is presented. This over-representation analysis is applied to get a further insight into biomass heterosis in early *Arabidopsis thaliana* development and to increase the confidence of identified candidate genes. Furthermore, the overlapping genes of the two approaches are tested for enrichment of *Arabidopsis thaliana* pathways from the TAIR and PO databases.
Finally, a general conclusion is given in chapter 5. Furthermore, in the appendix, supplementary analyses are presented. In appendix A the network hypothesis for heterosis, based on significances of the partial correlations, that was applied to the gene expression data in chapter 3, is tested on the metabolite data from chapter 2. This was not only done for mid-parent heterosis as in chapter 2 but also for best-parent heterosis. Appendix chapter B contains an analysis about the influence of the used cutoff value in the significance filtering step.

2 Towards Systems Biology of Heterosis: A Hypothesis about Molecular Network Structure Applied for the Arabidopsis Metabolome

Sandra Andorf[1], Tanja Gärtner[2], Matthias Steinfath[2], Hanna Witucka-Wall[3], Thomas Altmann[3], Dirk Repsilber[1]

[1]: Bioinformatics and Biomathematics Group, Genetics and Biometry Unit, Research Institute for the Biology of Farm Animals (FBN) Wilhelm-Stahl Allee 2, 18196 Dummerstorf, Germany
[2]: Institute for Biochemistry and Biology, University of Potsdam Karl-Liebknecht-Str. 24-25, 14476 Potsdam-Golm, Germany
[3]: Institute for Genetics, University of Potsdam Karl-Liebknecht-Str. 24-25, 14476 Potsdam-Golm, Germany

Published in *EURASIP J Bioinform Syst Biol*, articleID: 147157 (2009)

2.1 Abstract

We propose a network structure-based model for heterosis, and investigate it relying on metabolite profiles from *Arabidopsis*. A simple feed-forward two-layer network model (the Steinbuch matrix) is used in our conceptual approach. It allows for directly relating structural network properties with biological function. Interpreting heterosis as increased adaptability, our model predicts that the biological networks involved show increasing connectivity of regulatory interactions. A detailed analysis of metabolite profile data reveals that the increasing-connectivity prediction is true for graphical Gaussian models in our data from early development. This mirrors properties of observed heterotic *Arabidopsis* phenotypes. Furthermore, the model

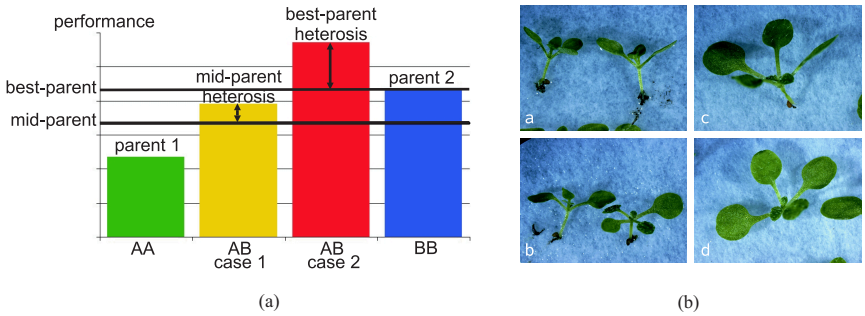
predicts a limit for increasing hybrid vigor with increasing heterozygosity – a known phenomenon in the literature.

2.2 Introduction

"Biological function" is the core of biological research, but it is an ill-defined term. Geneticists, cellular biologists, structural biologists, biophysical chemists and bioinformaticians all target different meanings in their respective research areas (Lambert and Hughes, 1984; Ge et al., 2003). However, as a unifying notion, biological function always refers to *semantic* features and, as such, is always context dependent. A specific state of any biological molecule alone is not accomplishing any biological function (Bohm, 1980). Rather, biological function resides in *interactions* (Strogatz, 2001; Somogyi and Sniegoski, 1996; Noble, 2002). The characteristics of such biological interactions, when analyzed on a genome-wide scale, are referred to as the *structure of biological networks* (including their dynamics). Relating structure of biological networks to biological function is therefore a major objective in biology, mirrored in recent developments such as systems biology.

A huge variety of biological networks exist, however, there are common characteristics: Biological network structure always arises as interaction of genetic determination and environmental influences, as well as internal systems dynamics. As pointed out by Somogyi and Sniegoski (1996), interactions within specific representations of biological networks may either map directly to existing biomolecules, or may reflect rather indirect relations involving possibly many of hidden variables (Perrot et al., 2007; Tresch and Markowetz, 2008). Most types of biological networks can be interpreted also as regulatory networks, in the sense that they "respond" to environmental or developmental challenges by changing their state or dynamics. A frequent approach to search for important network structures on a rather global level of biological networks is *statistical network modeling*. It starts out by screening for significant measures from graph theory (Barabási and Albert, 1999; Milo et al., 2002; Saul and Filkov, 2007). Distributions of such measures can then be compared between biological, technical or random networks, as well as between different classes of organisms (Milo et al., 2002; Lee et al., 2002; Matthäus et al., 2008), regimes of environmental challenges or developmental periods (Lee et al., 2002). If specific structures are discovered, their relation to a biological function of interest may be hypothesized and experimentally validated on further datasets.

In our case we are interested in contributing to a systems biological understanding of the biological phenomenon of heterosis. Shull (1908) defined the term heterosis as



(a) (b)

Figure 2.1: Definition of heterosis, (a): quantitative genetics definition of mid-parent heterosis and best-parent heterosis (heterosis effect: arrows); (b): example from early development in *Arabidopsis thaliana* – cotyledon areas are the largest in heterozygous crosses (c, d) as compared to their homozygous parents (a, b). [modified]

"increased vigor, size, fruitfulness, speed of development, resistance to disease and to insect pests, or to climatic rigors of any kind, manifested by crossbred organisms as compared with corresponding inbreds". See Figure 2.1a for a quantitative genetics definition of heterosis, and Figure 2.1b for an example of a trait showing a heterotic phenotype, cotyledon area in *Arabidopsis*. *Mid-parent heterosis* denotes an increase of performance relative to the mean of both parents, while *best-parent heterosis* describes the situation where the heterozygous offspring performs better than either parent. As early as in 1952, Robertson and Reeve (1952) suggested that heterozygotes are likely to posses a greater biochemical versatility by carrying a greater diversity of alleles. Heterosis would then result from a reduced sensitivity to environmental variations, since in heterozygotes there will be additional ways of overcoming such challenges. In other words, the heterosis phenomenon may be due to higher adaptability in heterozygotes. On the genetic level, hypotheses explaining heterosis may be grouped into two groups: On the one hand, dominant or overdominant modes of gene action are thought to play a major role, assuming recessive status for a majority of inferior alleles. On the other hand, enriched favorable epistatic interactions are discussed as main reason for the heterosis phenomenon at the molecular level (Birchler et al., 2003; Crow, 1952; Tsaftaris, 1995).

Gjuvsland et al. (2007) demonstrate how epistatic interactions within statistical genetics models can be translated into functional structures of regulatory biological networks. In our contribution we focus on these molecular network structures and ask: Which structures of biological networks could systematically lead to higher adaptability in heterozygotes, and, thus, to the heterosis phenomenon? For investigating this question we choose to follow a conceptual modeling approach (Somogyi

and Sniegoski, 1996; Wissel, 1992; Shubik, 1996). Our model choice is based on a major result of statistical network modeling. Analyses of distributions of simple regulatory motifs both in prokaryotes and in eukaryotes point to similar results: The so-called *multi-input-motif* is a significant and prominent part of regulatory biological networks (Milo et al., 2002; Lee et al., 2002; Shen-Orr et al., 2002). The properties of networks of this type were studied by Karl Steinbuch already in the year 1961 (Steinbuch, 1961). His studies were focusing on modeling and implementing models of associative learning. The so-called *Steinbuch matrix* is a two-layer feed-forward network. The information about which input vector is *associated* with which output vector is encoded within the pattern of presence/absence of connections between these two layers. We are going to use this Steinbuch network as a conceptual model for biological networks and develop a hypothesis of heterosis based on biological network structure. We expect specific global structures in biological networks to be different between homozygotes and their heterozygous offspring.

To validate and further detail our network hypothesis of heterosis we analyze partial correlation structures in experimental metabolite profile association networks from two different homozygous *Arabidopsis thaliana* lines and both reciprocal crosses as heterozygotes. These metabolite profiles were measured during early development of *Arabidopsis*, as it is during this time heterosis phenomena become manifest in this species (Meyer et al., 2004). We refer again to Somogyi and Sniegoski (1996) following their argument that not only the transcriptome, but also the metabolome could be viewed as a special mapping of the extended biological regulatory network. Such a mapping would include many indirect regulatory interactions involving hidden molecular variables which are part of other levels of gene expression.

Summarizing the objectives of our study, we motivate the proposal of a network-structure based hypothesis of heterosis and look for heterozygote specific network structures as predicted by a Steinbuch network conceptual modeling approach. Analyses of metabolite profiles of early development in *Arabidopsis thaliana* and further observations of heterosis in plants will serve as to validate and further adjust our hypothesis.

Section 2.3 describes the experimental dataset and our pre-processing prior to statistical network analyses. In section 2.4 we describe our modeling approach as well as a small simulation study. Its results motivated our choice of network statistics for global assessment of network structures described in the remaining part of this section. The first part of section 2.5 reports the simulation results. In its second part we develop our network-structure based hypothesis of heterosis and its predictions. In the last part of this section results of experimental data analysis as motivated by

our model's predictions are presented. Finally, in section 2.6 we discuss the main findings of our study, their relevance, benefits and constraints of our approach as well as future prospects.

2.3 Experimental Data and Preprocessing

We investigate metabolite profiles, GC-MS data, of early development of *Arabidopsis thaliana*. More precisely, metabolite profiles of plants of the two homozygous lines C24 and Columbia (Col-0) and the reciprocal crosses, Col-0 × C24 and C24 × Col-0, are studied. Metabolite profiles of the two homozygous genotypes, C24 × C24 and Col-0 × Col-0, and the two heterozygous genotypes, C24 × Col-0 and Col-0 × C24, were measured at 7 time points (0, 12, 24, 36, 48, 72, 96 hours after sowing (HAS)). For each measurement a petri-dish of seedlings was grown and fully harvested after the specific time of growing. In our balanced cross-factorial design, four replicates were assessed per genotype and time point, measured at three different measuring days, such that each genotype-time point combination was measured at least once per measuring day. The raw data preparation was performed as in Lisec et al. (2006), afterwards, the data were log-transformed. Overall 210 metabolites have been measured. Eight of them contained more than 20% missing values and were therefore excluded from further analysis.
For normalization we chose a linear modeling approach, involving the factors $g \in \{\text{C24} \times \text{C24}, \text{Col-0} \times \text{Col-0}, \text{C24} \times \text{Col-0}, \text{Col-0} \times \text{C24}\}$ denoting the four genotypes, factor $t \in \{1, ..., 7\}$ denoting the 7 time points of the developmental time series, their interaction $g \times t$, as well as factor $d \in \{1, ..., 3\}$ denoting the measuring day. The linear regression was fit on a per metabolite basis for the following model, for which y, the logarithm of the raw metabolite signal, is modeled as dependent on the factors described above:

$$y_{i,j,k,l} = \mu + g_i + t_j + (g \times t)_{i,j} + d_k + \varepsilon_{i,j,k,l}. \tag{2.1}$$

Here, μ gives the overall mean, the four genotypes are denoted with index i, the seven time points with index j, the measuring days with index k and the replicates with index l. Normalized metabolite profiles were obtained using the effect estimates from the fit of model 2.1 as in Eqn. 2.2. This way, data were corrected for measuring day effects as well as correct mean values were calculated, even for combinations with single missing values.

$$y^*_{i,j} = g_i + t_j + (g \times t)_{i,j}. \tag{2.2}$$

The resulting time series of normalized metabolite profiles is plotted in Figure 2.2 for genotype C24 × C24.

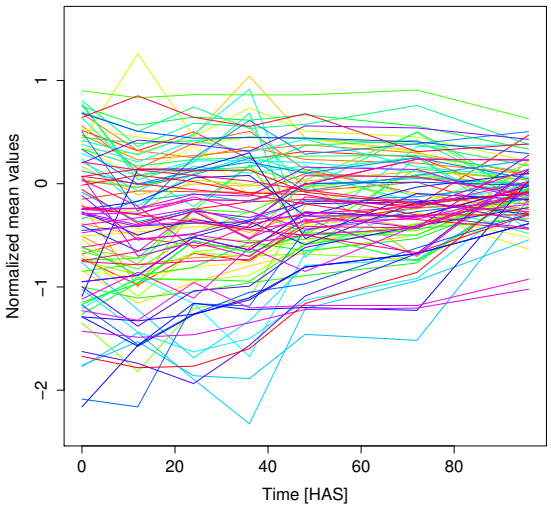


Figure 2.2: Profiles of normalized values for each metabolite (202 different colors) over seven points for the genotype C24 × C24 as obtained from Eqn. 2.2.

2.4 Methods

2.4.1 Modeling and Simulation

Our conceptual modeling approach employs a model of association to simulate adaptability in regulatory networks: Adaptedness can be described as the ability to give a correct response (output) to an environmental or developmental challenge (input). Hence, an adaptation can be viewed as the correct *association* of a response to the input in question. Correspondingly, adaptability is the number of differentiated correct adaptations a regulatory system is able to realize.

Figure 2.3a shows a scheme representing a diploid genome and various levels of gene expression (transcriptome, proteome, metabolome). Black arrows represent *synthesis*, colored arrows symbolize *regulatory functions*. Simplifying this scheme leads to the simplest possible homomorphic model, an association matrix as in Figure 2.3b. Here, input and output are associated via the interactions between input layer and output layer. In the output layer, signals from the input layer are summed up and compared to a threshold cutoff as to yield an output of "1" if larger or equal, or of "0" if smaller. The association network can be modeled mathematically as an

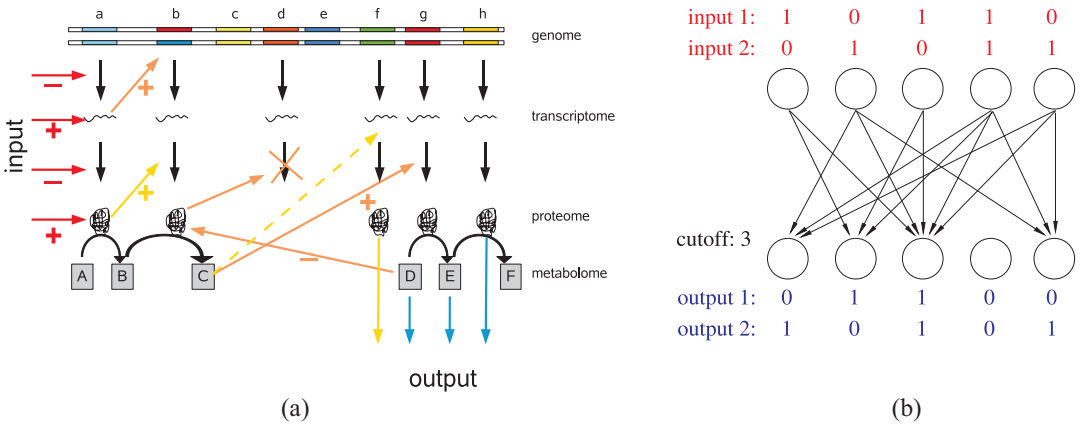


Figure 2.3: Schematic representation of molecular networks (a) with *synthesis* (black arrows) and *regulatory* functions (colored arrows), as homomorphic to the association network model (b), representing a two-layered feed-forward Steinbuch matrix: Associated input-output pairs are depicted in corresponding colors (blue and red). Black arrows depict regulatory interactions between specific input and output nodes.

n x n matrix, $\mathbf{R}$, where n denotes the size of the network which is given by the number of nodes in the input and output layer, respectively (e.g. $n = 5$ for the network in Figure 2.3b). In this model each molecular entity (metabolite, protein, transcript) has two possible states, "0" or "1". The input signal, s_{in}, is converted into the output s_{out} through

$$s_{out} = \theta(\mathbf{R} \cdot s_{in}) \tag{2.3}$$

where θ is a threshold function that is applied component wise:

$$\theta([\mathbf{R} \cdot s_{in}]_i) := \begin{cases} 1 & \text{if } [\mathbf{R} \cdot s_{in}]_i \geq \vartheta_i \\ 0 & \text{if } [\mathbf{R} \cdot s_{in}]_i < \vartheta_i \end{cases} \tag{2.4}$$

where for example $\vartheta_i = max_i([\mathbf{R} \cdot s_{in}]_i)$.
For the case given in Figure 2.3b, the matrix for the association network is given by

$$\mathbf{R} = \begin{pmatrix} 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 \end{pmatrix}. \tag{2.5}$$

We conducted a small simulation study, employing an association matrix of size

$n = 150$ which is capable of correctly associating 4 pairs of input-output vectors. The model was trained to reproduce these pre-defined input-output pairs, which can be interpreted as some kind of crucial regulatory reply (regulatory step) to cope with a special environmental challenge. The study should reveal whether a partial correlation analysis of state profiles for the nodes of the network is a valid possibility to study the causal regulatory interactions in this network. 100 randomly generated input vectors, s_{in}, and their corresponding outputs, s_{out}, were stored as profile data and partial correlations calculated as detailed in what follows.

2.4.2 Network Statistics

Different types of networks can be used to assess the underlying biochemical interaction network from high-throughput metabolomic data. For our analysis we have used partial correlations. This belonging network is known as graphical Gaussian model (GGM), concentration graph, covariance selection graph, conditional independent graph (CIG) or Markov random field (Opgen-Rhein and Strimmer, 2007b). Partial correlations have been shown to be a suitable method for deducing regulatory interactions from observational (non-interventional) data (Werhli et al., 2006). They are calculated by Opgen-Rhein and Strimmer (2007b) from metabolite levels as in Eqn. 2.6.

$$\tilde{\rho}_{k,l} = \frac{-\omega_{kl}}{\sqrt{\omega_{kk}\omega_{ll}}} \tag{2.6}$$

The basis for these values are the normalized metabolite values for the seven time points from Eqn. 2.2 for each genotype and each of the analyzed 202 metabolites. Thus, for any two metabolites of one of the four genotypes, partial correlations can be calculated based on the seven pairs of metabolite values corresponding to the seven time points. $\tilde{\rho}_{kl}$ is the estimate of the partial correlation between the metabolites k and l. ω are the elements of the inverse covariance matrix which is estimated using a shrinkage estimator (Schäfer and Strimmer, 2005b). The algorithm is implemented in the *R* package *GeneNet* (Opgen-Rhein et al., 2007).
We investigate changes for the partial correlation structure between heterozygous and homozygous genotypes by first calculating a "mid-parent"-value as mean value for each metabolite and both homozygous genotypes:

$$\tilde{\rho}_{m,n}^{\,midparent} = \frac{1}{2} \sum_{i \in \{\text{C24}\times\text{C24,Col-0}\times\text{Col-0}\}} \tilde{\rho}_{i,m,n} \tag{2.7}$$

for all metabolites $m, n \in \{1, ..., 202\}$.

Second, the heterosis effects were calculated for both heterozygotes as increase of absolute partial correlation in the heterozygote compared to the mid-parent value. These values were calculated for all pairwise combinations of metabolites (Eqn. 2.8, compare Figure 2.1a). We considered absolute correlations because an increase of positive correlations should be equally weighted as a decrease of a negative correlation:

$$\tilde{\rho}_{k,m,n}^{\ heterosis} = |\tilde{\rho}_{k,m,n}| - |\tilde{\rho}_{m,n}^{\ midparent}|. \tag{2.8}$$

Here, k denotes the respective heterozygous line ($k \in \{\text{C24} \times \text{Col-0}, \text{Col-0} \times \text{C24}\}$).

Third, to characterize changes in partial correlation with respect to the mid-parent value on a per-metabolite basis, for each metabolite met $\in \{1, ..., 202\}$ we calculated the mean values across all pairs involving this metabolite[1]:

$$\tilde{\bar{\rho}}_{k,\text{met}}^{\ heterosis} = \frac{1}{201} \sum_{l \in \{1,...,202\}, \text{met} \neq l} \tilde{\rho}_{k,\text{met},l}^{\ heterosis} \tag{2.9}$$

Distributions of $\tilde{\bar{\rho}}_{k,\text{met}}^{\ heterosis}$ were displayed and compared.

To investigate if the metabolites showing the largest values for $\tilde{\bar{\rho}}_{k,\text{met}}^{\ heterosis}$ had a specific distribution over metabolite pathways we visualized the first thirty metabolites in a ranking of $\tilde{\bar{\rho}}_{k,\text{met}}^{\ heterosis}$ for each heterozygous line using MapMan (Thimm et al., 2004). MapMan is a tool to display large datasets onto diagrams of metabolic pathways.

Not only global distributions of changes in partial correlations could be different between homozygous and heterozygous lines, but also structural properties of partial correlation networks. In such networks *edges* are significant partial correlations, computed according to Opgen-Rhein and Strimmer (2007b). P-values were corrected using the FDR correction described by Benjamini and Hochberg (1995). Accordingly, *nodes* in partial correlation networks are the metabolites contributing to significant partial correlations.

The *degree* of such a node is defined as the number of edges it is part of. We characterized the partial correlation networks of the two homozygous and the two heterozygous lines by counting significant edges and the participating nodes, as well as calculating the mean degree values over all nodes of a network.

[1] formula slightly modified

2.5 Results

2.5.1 Simulation Results

When comparing association matrices capable of reproducing an increasing number of associations ($p \in \{1 \dots 4\}$), the belonging networks show an increasing number of causal interactions between input and output layer (see Figure 2.4a).
Our small simulation study, where we recorded outputs for 100 random inputs to a 150x150 association matrix reproducing 4 input-output associations, revealed that causal interactions between input and output layer lead to increased partial correlations of the respective nodes.
As demonstrated in Figure 2.4b, for our model causal interactions can be deduced from observational profile data by calculating partial correlations. These properties of our conceptual model led to the development of a network-structure based model of heterosis as outlined in what follows.

2.5.2 Network Hypothesis of Heterosis

As suggested by Robertson and Reeve (1952) heterozygotes are likely to posses a greater biochemical versatility by carrying a greater diversity of alleles. Heterosis would then result from a reduced sensitivity to environmental variations, since there will be ways of overcoming such challenges. In other words, the heterosis phenomenon may be due to higher adaptability in heterozygotes.
Correspondingly, as illustrated in Figure 2.3a, the molecular network of a heterozygous cross may contain a proportion of heterozygous loci, as for gene "b" for example. The additional alleles at this locus may lead to *additional* regulatory interactions in the molecular network (yellow arrows in Figure 2.3a). In our model, as shown in the simulation (see Figure 2.4b), additional causal interactions are the basis of an increasing number of associations in the repertoire of the Steinbuch network.
It is known from earlier studies of system properties of the Steinbuch network that there exists a *limit of associated pairs* for a network of a given size (Nadal, 1991). A Steinbuch network of a given size can be build to be able to differentiate between a certain number of inputs by "responding" with the (associated) belonging outputs, and not more. This is a known system property of this type of regulatory network – but also for other types of neural networks.
Moreover, if we measure an increasing amount of partial correlations within a molecular network, this might correlate with an increased amount of regulatory "challenge-response" pairs managed by this network, and, hence, with increased adaptability.

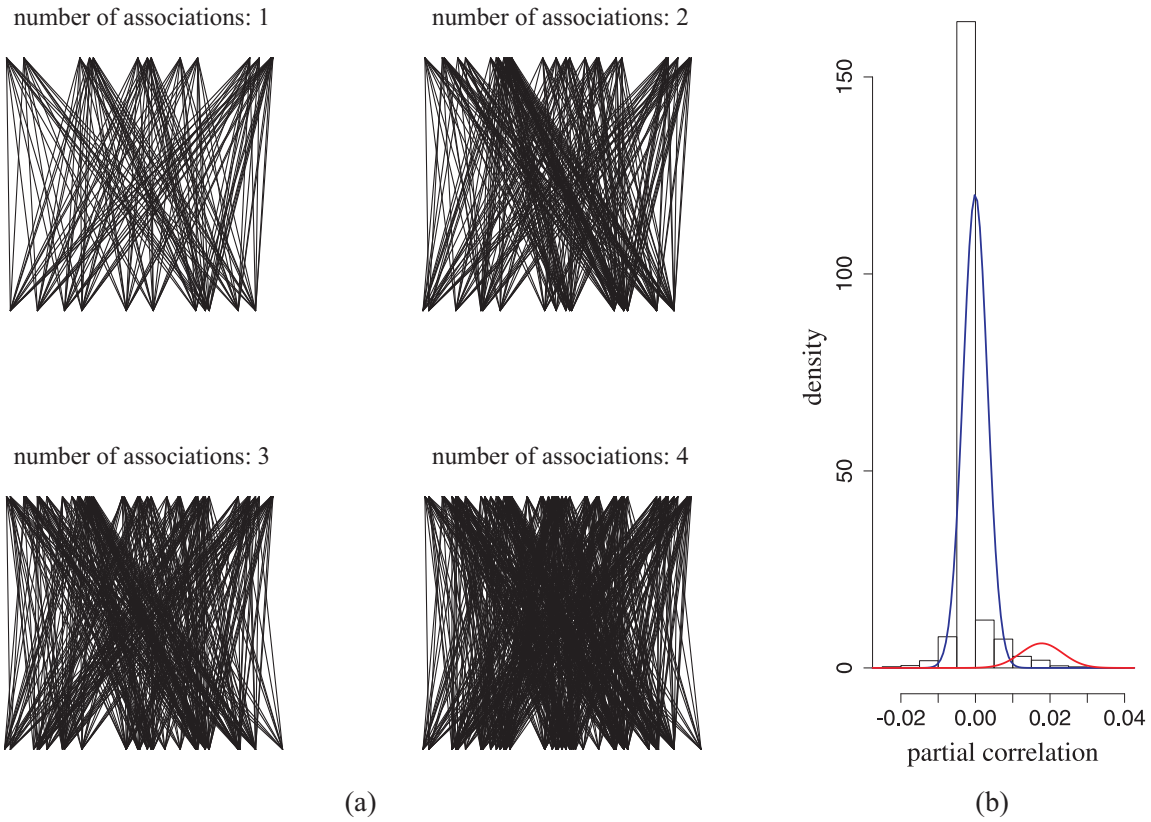


Figure 2.4: Example for a 150x150 Steinbuch matrix. (a): Increase in number of regulatory interactions between input and output layers, representing an increasing number of association pairs. (b): Analysis of the matrix of (a) with the ability to reproduce 4 pre-defined association pairs. Distribution of partial correlations for non-interacting input-output nodes (blue, entry "0" in R) and for interacting input-output nodes (red, entry "1" in R).

Interpreting these properties as conceptual model for adaptation and adaptability in molecular regulatory networks leads to two predictions for the case of heterosis:

1. There should exist a limit for increasing hybrid vigor with increasing level of heterozygosity. Increasing the genetic distance of homozygous parental lines beyond a certain threshold should result in less hybrid vigor if these parental lines are genetically too different. When mating two similar homozygous genotypes, only few additional regulatory connections within the molecular networks can be expected. However, when mating homozygous genotypes which are genetically very different (with large genetic distance) the limit of the resulting merged molecular network structures may be exceeded – in the sense that regulatory interactions in the network of the resulting heterozygotes do not match and therefore do not lead to additional possibilities of valid regulatory answers.

2. Molecular interactions in regulatory networks of heterozygotes should be slightly enriched. This increased number of "challenge-response"-pairs are modeled as a higher number of association pairs in our conceptual model, interpretable as increased adaptability leading to heterosis. As for the model, where we were able to measure interactions as increased partial correlations, we also expect an increase in partial correlations from homozygotes to heterozygotes for the experimentally observed dynamics of biological regulatory networks.

For evaluating prediction 1 we had no own experimental data, as these were only based on crosses of two homozygous lines. Instead, we analyzed the literature basis of a possible relationship between heterosis and genetic diversity. Figure 2.5 summarizes this literature view regarding a possible limit of gain in hybrid vigor in offspring for

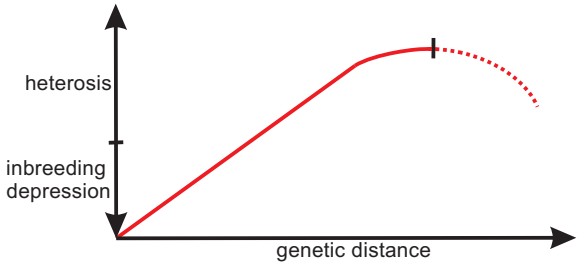


Figure 2.5: Possible relationship between genetic distance of the parental lines and hybrid vigor in the offspring. There is evidence for the existence of a limit of increase in hybrid vigor, as indicated in Moll et al. (1965); Link et al. (1996); Melchinger (1999) and Falconer and Mackay (1996).

increasing genetic diversity between the parental lines. From studies in maize as well as beans it seems likely that, with increasing genetic diversity between the parental lines, resulting hybrid vigor for the offspring at first increases. However, for parental lines which are genetically too different it is expected to decrease again (Moll et al., 1965; Link et al., 1996; Melchinger, 1999; Falconer and Mackay, 1996). We want to emphasize that, given the literature basis as investigated, further research on the first part of our network hypothesis of heterosis seems promising – and necessary, as at the moment we cannot draw stronger conclusions.

Regarding prediction 2 we studied our experimental dataset, the *Arabidopsis* metabolome of a developmental time series (see section 2.5.3). From the perspective of our model, Figure 2.3a illustrates how the molecular network of heterozygotes contains additional regulatory possibilities. In the association network model these correspond to additional connections (interactions) between input and output layer, enabling the network to add additional associations to its repertoire. These additional associations (input-output-pairs) represent a grown repertoire of adaptations, or increased adaptability, enabling increased hybrid vigor. The objective of our experimental data analyses was to investigate if such increase in molecular interactions would be measurable as increase in partial correlations as a global network property for the metabolite profiles recorded during *Arabidopsis* development.

2.5.3 Analysis of Experimental Data

Our experimental data were metabolite profiles from development of *Arabidopsis thaliana* (see Figure 2.2). To test our hypothesis that heterosis comes as increasing adaptability and should result in increasing connectivity of molecular networks, we had first conducted a small simulation study (see section 2.5.1). Its findings provide the basis for our investigation of partial correlation structures of the metabolomes of heterozygous and homozygous genotypes for the experimental data, as we want to test a hypothesis about increased regulatory possibilities in heterozygotes and the belonging structures of molecular profiles. Hence, we now analyzed partial correlations according to Opgen-Rhein and Strimmer (2007b) for our experimental dataset.

The average heterosis increase of the partial correlations in the heterozygous lines as compared to the mid-parent value (mean of the homozygous lines) was calculated ($\tilde{\tilde{\rho}}^{\,heterosis}_{k,\mathrm{met}}$, see Eqn. 2.9). Results are displayed in Figure 2.6. The histograms for $\tilde{\tilde{\rho}}^{\,heterosis}_{\mathrm{C24\times Col\text{-}0,met}}$ for the genotype C24 × Col-0 (Figure 2.6a) as well as $\tilde{\tilde{\rho}}^{\,heterosis}_{\mathrm{Col\text{-}0\times C24,met}}$ for the genotype Col-0 × C24 (Figure 2.6b) show that for a majority of the metabolites the calculated difference is positive. That means that the mean partial correlation

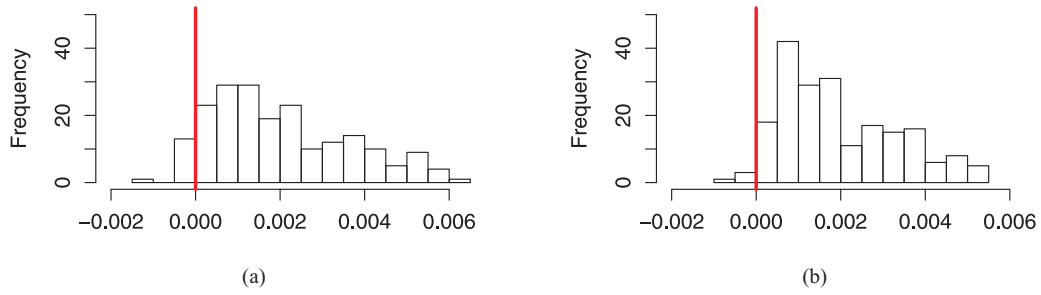


Figure 2.6: Display of $\tilde{\tilde{\rho}}_{k,\text{met}}^{\ heterosis}$ for $k \in \{\text{C24} \times \text{Col-0}, \text{Col-0} \times \text{C24}\}$ (see Eqn. 2.9). The mean differences for the most metabolites between the partial correlations of genotype C24 × Col-0 (a) as well as Col-0 × C24 (b) to the average of the homozygotes (mid-parent) are positive values.

values of either heterozygous genotype are larger than the average of the homozygotes (mid-parent).

For each heterozygous genotype the 30 metabolites that show the largest difference were determined. For the genotype C24 × Col-0 these selected metabolites are displayed onto a diagram of biochemical pathways in Figure 2.7 using MapMan (Thimm et al., 2004) to study possible pathway-related differences in the partial correlation values between homozygous and heterozygous genotypes. Metabolites of the top 30 are marked as red points. The picture does not contain 30 red points because the top 30 list contains several unknown metabolites. Furthermore, not all metabolites are available in the MapMan annotation. The displayed metabolites are relatively evenly distributed over all illustrated pathways. For the genotype Col-0 × C24 this distribution looks similar (data not shown). 12 metabolites were in common for the top 30 lists of both heterozygous genotypes.

Table 2.1: Significant partial correlations (significance level $\alpha_{FDR} = 0.1$).

Genotype	No. sign. edges	Corresp. nodes	Mean degree
C24 × C24	10	13	1.54
Col-0 × Col-0	23	23	2.00
C24 × Col-0	81	45	3.60
Col-0 × C24	64	40	3.20

In Table 2.1 the detailed results of the connectivity analysis are listed. For all metabolites the partial correlations are based on the time series of the 7 time points from 0 HAS to 96 HAS. In the table, the number of significant edges and the number

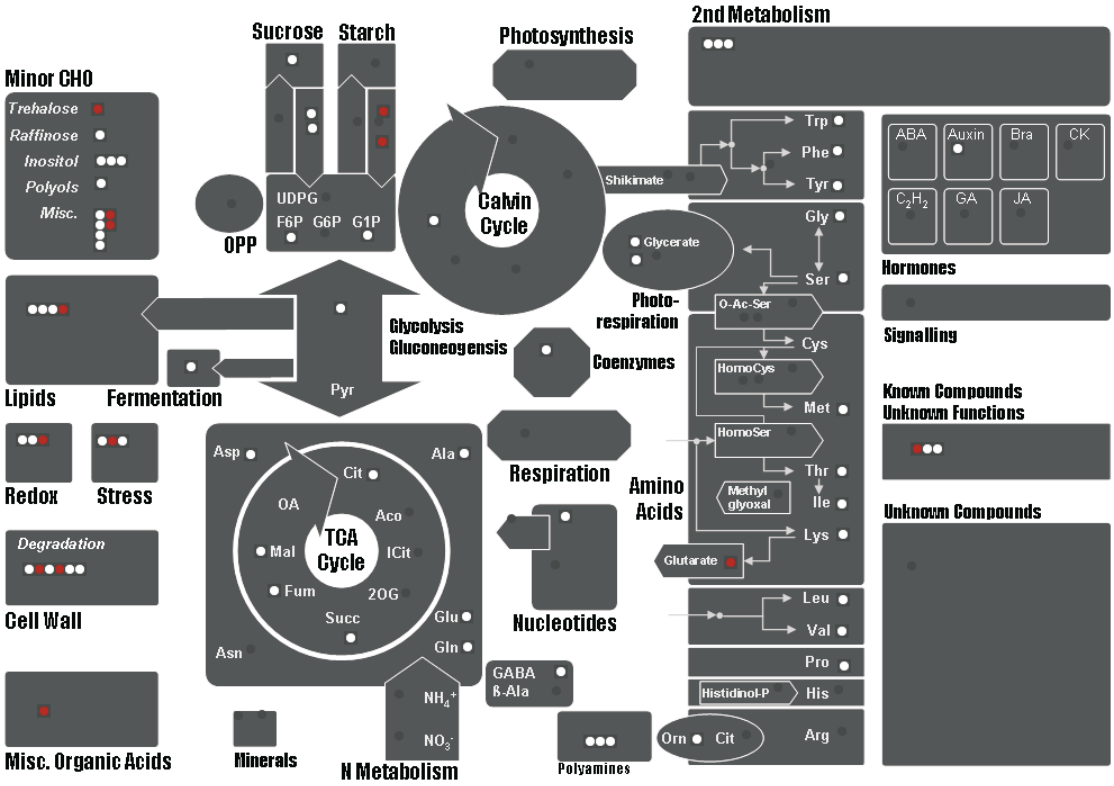


Figure 2.7: Metabolites with highest mean differences between absolute partial correlation values of genotype C24 × Col-0 and the mean of the homozygous lines are displayed on plant biochemical pathways (red). White: metabolites that are present in the MapMan (Thimm et al., 2004) annotation list as well as in our metabolite list but not within the top 30 list. Dark gray: not measured.

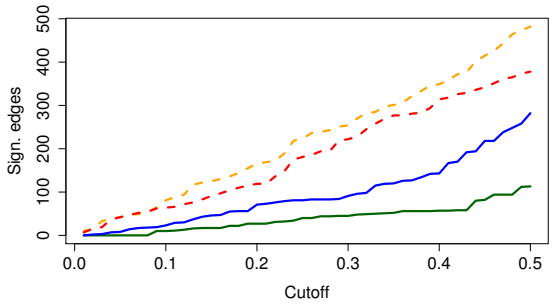


Figure 2.8: Display of numbers of detected significant partial correlations as dependent on corrected P-value cutoff (significant partial correlations) for the 4 genotypes. Heterozygotes (dashed lines) show a higher number of significant edges throughout. (C24 × C24: green, Col-0 × Col-0: blue, C24 × Col-0: orange, Col-0 × C24: red)

of nodes (metabolites) that belong to these edges are shown. Our main focus in this analysis was on mean degree. These mean degree values were calculated on the basis of the number of nodes with significant edges (see definition at the end of section 2.4.2).

Both homozygous genotypes show lower mean degrees than either heterozygote. As shown in Figure 2.8 the relation between the number of significant edges of the heterozygotes and those of the homozygotes is nearly independent from the cutoff used.

We choose a cutoff $\alpha_{FDR} = 0.1$ for the FDR-corrected P-value to determine the significant edges in each analysis. This outcome is illustrated in Figure 2.9: The partial correlation networks of the two heterozygous genotypes show more connections than the networks of the homozygous genotypes.

Hence, results of Figure 2.6 and Figure 2.9 point towards the same tendency, supporting the "increased-connectivity"-prediction of our network hypothesis of heterosis. This tendency is strengthened as most of the 30 metabolites that show the largest differences between the heterozygotes and the mid-parent value also have significant edges. In more detail, for genotype C24 × Col-0, 25 of the top 30 metabolites and, for genotype Col-0 × C24, 27 of the top 30 metabolites have significant edges. Total numbers of nodes with significant edges are 45 and 40 respectively (see Table 2.1). In average, for either heterozygous genotype 86.7% of the top 30 metabolites show significant edges.

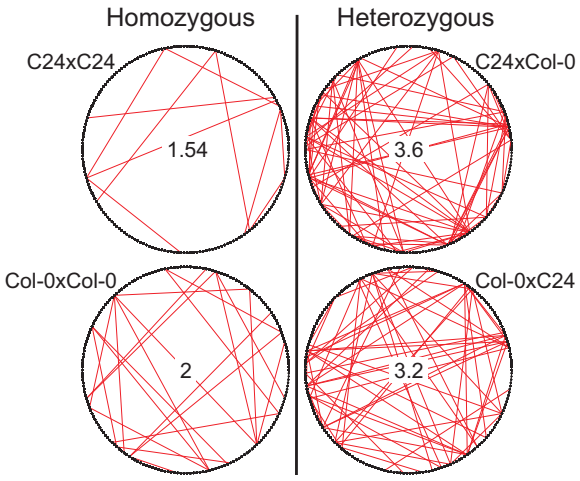


Figure 2.9: Connection plots based on partial correlations, using a cutoff $\alpha_{FDR} = 0.1$ for the belonging FDR-corrected P-values. The heterozygous genotypes show more significant edges and a higher connectivity than the homozygous genotypes. Mean degrees are given for each genotype.

2.6 Discussion

We have developed a network structure-based hypothesis of heterosis. It is a systems biological approach to relate biological function to molecular network structure. Our hypothesis results in the following predictions: First, system properties of our network modeling approach suggest the existence of an upper limit for the heterosis effect when genetic distance of crossed homozygous parental lines becomes too large. Second, molecular networks of heterozygotes should contain additional interactions compared to those of their homozygous parents. These additional interactions should lead to increased partial correlations in molecular networks of heterozygotes. For the first prediction, we found support in the literature suggesting an upper limit for the heterosis effect. However, as we do not have sufficient additional own experimental evidence, no final conclusion can be drawn for this case. Further investigations seem promising and necessary. Regarding the prediction of increased connectivity of molecular networks in heterozygotes, for our own experimental metabolome dataset of *Arabidopsis* such increased connectivity was observable for both heterozygous crosses. It is this phase of early *Arabidopsis* development in which the heterosis effect is established. The predicted pattern is visible for the majority of metabolites. However, also for the second part of our network hypothesis of heterosis, we call for additional experimental evidence, preferably on additional levels of molecular regulatory networks, such as proteomics or transcriptome data. Summarizing, we present a conceptual frame for explaining the heterosis phenomenon from a molecular network perspective together with two hypotheses and their predictions, for which we were able to find first supporting evidences from the literature and own experimental data.

We are convinced that research towards understanding the biological phenomenon of heterosis can particularly gain from a systems biological approach focused on *interactions* of molecular building blocks and global structures of molecular biological networks. Towards elucidating the genetic basis of heterosis, Melchinger et al. (2007b) have already shown that, taking a statistical modeling approach, epistatic interactions of individual loci with the entire genetical background constitute a major component of genetic variation important to explain heterosis. However, the mapping between interaction terms in models of quantitative genetics to structures in molecular regulatory networks is non-trivial (Gjuvsland et al., 2007; Gibson, 1996). Our approach to investigate global network structures in molecular interaction networks for this reason is to be taken as *complementary* to the quantitative genetics view.

Meyer et al. (2004) report for *Arabidopsis thaliana* development, that it is the *early* phase of development (until one week of seedlings growth) during which the heterosis phenotype for biomass is established. In later phases of the plant's life relative differences between heterozygotes and homozygotes are not further growing. The first observation coincides with our results: We observe increased connectivity in partial correlation networks in this period of development. It would be interesting to see – and is planned as future experimental study – if during the later phase, when according to Meyer et al. (2004) biomass heterosis is visible but no longer increasing, there is no indication of increased connectivity in the metabolome any longer.

The majority of metabolites investigated showed an increase in interaction connectivities. We tried to find common functionalities for the top thirty metabolites with most obvious changes. However, we were not able to detect evidence towards an accumulation of such metabolites within certain pathways or modules (MapMan categories). We hypothesize that it may be these metabolites which during the early phase in *Arabidopsis* development are mainly involved in *regulatory interactions* – to enable adaptation to the climatic chamber during the first contact with this environment.

Only part of the observed changes in partial correlations between heterozygous lines and the mid-parent value of both homozygotes can be based upon *significant* partial correlations (compare Figure 2.6 and Figure 2.9). However, the same tendency is apparent for the global view as well as for the restriction to significant correlations. It is the sparsely designed experimental data which does not allow a more precise analysis. Seven time points are clearly the *lower limit* of correlation analyses involving around two hundred metabolite species. We look forward to more generously designed experiments for testing our network-structure based hypotheses for heterosis.

Our modeling approach is *conceptual* as advocated for by, for example, Wissel (1992) and Shubik (1996). It builds upon the understanding of the heterosis phenomenon as increased adaptability. This understanding has its roots already at the beginning of the 20th century in maize genetics (Shull, 1908) and since then has been expressed also within the context of hybrid vigor observed for other plant species as well as model animals (see, e.g., Robertson and Reeve (1952); Solomon et al. (2007); Harrison (1962)). We make use of a model for adaptability which was originally designed to model associative memory, the Steinbuch matrix (Steinbuch, 1961). Within our model, being *adapted* means to respond in a correct way when confronted with a certain environmental or developmental stimulus – while *adaptability* means the potential to respond to a number of different stimuli with differentiated cor-

rect responses. The simplicity of this conceptual modeling implies rather *general* predictions. In our case these are the limit-of-heterosis-increase prediction and the increasing-connectivity prediction. These are predicted for a huge class of interaction networks, independent of molecular species. Motif analyses in different molecular interaction networks as well as within organisms of different kingdoms (prokaryotes, eukaryotes) have shown that certain motifs are always present. The "multi-input-motif" is a prominent example. Here we refer to the work by Milo et al. (2002) and Lee et al. (2002). The *multi-input-motif* has the same structure as our association network model, which was first proposed already 1961 by Steinbuch (1961). Furthermore, molecular interactions are often modeled based on a sigmoidal relationship as approximated by the boolean kind of interaction in the Steinbuch model (discussed in Kauffman, 1993).

A central assumption underlying motif analyses as well as our modeling approach for this work is that neglecting the diversity of different kinds of molecular species, that interact within real molecular networks, does not harm at the rather general level of conclusions of our conceptual investigations. It is clear that natural molecular networks cannot be reduced to a very simplistic model in *all* their structural and dynamical properties. However, we chose to follow Shubik's call for the most *parsimonious* modeling approach (Shubik, 1996). Also, heterosis is a very general biological phenomenon together with its counterpart inbreeding depression. Both phenomena are occurring over a broad variety of sexually reproducing organisms. For this reason, approaches towards understanding the systems biological foundations of these phenomena should be independent of all organism specific parameters, in other words as simple as possible.

Choosing the metabolome level, as in our study, is just one possibility. With Somogyi and Sniegoski (1996) we argue that the *extended regulatory network* of an organism can be mapped to any of its levels of gene expression ("omics" levels). However, the modeler has to be aware of all possible hidden variables constituting each of the investigated interactions. These hidden variables are representations of the molecules from the "omics" levels which were not modeled. In our case for example, regulatory interactions between metabolites have no direct correspondence to metabolic pathways. Moreover, as is true for gene expression studies for the case of transcription factors, also in metabolomics it is not at all possible to assess *all* molecules, but only a small fraction. The measurable fraction may or may not be a biased sample from the entire metabolome – and for this reason inferring network structures from such a sample has always to be taken with care (for an example concerning network statistics in protein interaction networks see de Silva et al. (2006)). Also, we are aware

of the problem of cell type heterogeneity in our samples which are basically whole embryo/plant homogenates. Measured profiles in our case represent metabolite levels of the major cell type. In addition, it is important to take into account the fact that those 202 metabolites in our investigation are just around 10% (possibly less) of the metabolites that are suposed to be present in *Arabidopsis thaliana* (Cui et al., 2008). Thus, our network structure-based hypothesis of heterosis was validated only for the core carbon metabolism. These small molecules, as for example sugars, amino acids and carbon acids, act mostly within energy metabolism and as precursors for building the larger biomolecules, proteins, nucleic and fatty acids. These metabolites represent what is currently measurable with the GC-MS metabolite profiling experiments.

For future investigations of molecular network structures with respect to the heterosis phenomenon it will be an interesting challenge to extend the time series design of the current study in several aspects. To enable a more general conclusion regarding the two predictions from our network hypothesis of heterosis it would be worth comparing several different homozygous lines and their reciprocal offspring. Also, genetically very different lines should be included to approach a direct test of the limit-of-heterosis-increase prediction. Moreover, time points should be set more dense, e.g. as 10 hours intervals, and over a longer time-scale, e.g. at least along the first four weeks of *Arabidopsis thaliana* development. Such a design would enable both a higher precision for estimating partial correlation structures, as well as assessing a possible change of such structures during later phases of development – for which according to Meyer et al. (2004) no additional heterosis effects are arising. Furthermore, studies are already planned to analyze *transcript data* measured under the same conditions as our metabolome dataset. This would enable to show, first, how two levels of the extended regulatory network act together taking an integrative bioinformatics approach (see for example Steinfath et al. (2007)). Second, it would be possible to test the increasing-connectivity prediction of heterosis also for the level of the transcriptome.

Regarding alternative approaches to measure differential network structures in molecular networks of homozygotes and heterozygotes there exist a number of possible choices. An alternative type of networks used for inference of biochemical interaction networks are for example the so-called relevance networks. Butte et al. (2000) base their method on a pairwise Pearson correlation of all features. A serious limitation of relevance networks is, that they contain many indirect correlations, because they cannot distinguish between direct and indirect interactions. For our kind of *observational* data Werhli at al. have shown that it is preferable to use association networks

to infer regulatory interactions (Werhli et al., 2006). For this reason, we decided to analyze partial correlations as proposed by Opgen-Rhein and Strimmer (2007b). We also favored the regularized inference of the covariance matrix they proposed, which is applicable for data with a small sample size and a comparatively large number of variables, as in our metabolome dataset. Our simulation study was able to demonstrate that, when observing a number of partial correlations from the Steinbuch model, these could be used to identify the nodes of input and output layer which were connected in the regulatory architecture of the network model to reproduce four pre-defined input-output patterns. Hence, for our conceptual model, regulatory interactions could be deduced from partial correlations. A possibly promising way to extend our analyses could be oriented along the lines of the work by Saul and Filkov (2007) who proposed to use so-called exponential random graph models. They demonstrate their utility in modeling the architecture of biological networks as a function of a number of different measures of local network structure, not only a single measure as in our case. The flexibility, in terms of the number of available local feature choices, and scalability possibly make this approach a suitable alternative for statistical modeling of biological networks.

To summarize, in our work we followed the call of Barabási and Oltvai (2004) who conclude their review on *network biology* by stating that structure, topology, network usage, robustness and function are deeply interlinked, forcing us to complement the 'local' molecule-based research with integrated approaches that address the properties of regulatory networks on a systems biological level. In our study we have done so, by proposing a network structure-based model of heterosis and investigating its predictions for an experimental omics-dataset: Heterotic phenotypes of *Arabidopsis* are mirrored as increased connectivity in metabolome partial correlation networks. A limit of hybrid vigor increase for increasing genetic distance of crossed parents is also correctly predicted. These results hold for the measured part of the metabolome, mostly central carbon metabolism.

Our conclusions cannot be more than an illustrative example of how a hypothesis can be built about a possible relation of biological network structure to biological function, in our case the heterosis phenomenon. We advertise our approach as a way of investigating heterosis complementary to the quantitative genetics approach and look forward to future unifying approaches to these two fields.

Acknowledgments

This work was supported by DFG under grants RE 1654/2-1 and SE 611/3-1.

3 Enriched partial correlations in genome-wide gene expression profiles of hybrids (A. thaliana): A systems biological approach towards the molecular basis of heterosis

Sandra Andorf[1], Joachim Selbig[2], Thomas Altmann[3], Kathrin Poos[4], Hanna Witucka-Wall[2], Dirk Repsilber[1]

[1]: Research Institute for the Biology of Farm Animals (FBN)
Wilhelm-Stahl Allee 2, D – 18196 Dummerstorf, Germany
[2]: University of Potsdam
Karl-Liebknecht-Str. 24-25, D – 14476 Potsdam-Golm, Germany
[3]: Leibniz Institute of Plant Genetics and Crop Plant Research (IPK)
Corrensstr. 3, D – 06466 Gatersleben, Germany
[4]: University of Applied Sciences Gelsenkirchen, Site Recklinghausen
August-Schmidt-Ring 10, D – 45665 Recklinghausen, Germany

Published in *Theor Appl Genet*, 120:249–259 (2010)
(With kind permission from Springer Science and Business Media.)

3.1 Abstract

Heterosis is a well-known phenomenon but the underlying molecular mechanisms are not yet established. To contribute to the understanding of heterosis at the molecular level, we analyzed genome-wide gene expression profile data of *Arabidopsis thaliana* in a systems biological approach. We used partial correlations to estimate the global interaction structure of regulatory networks. Our hypothesis states that heterosis

comes with an increased number of significant partial correlations which we interpret as increased numbers of regulatory interactions leading to enlarged adaptability of the hybrids. This hypothesis is true for mid-parent heterosis for our dataset of gene expression in two homozygous parental lines and their reciprocal crosses. For the case of best-parent heterosis just one hybrid is significant regarding our hypothesis based on a resampling analysis. Summarizing, both metabolome and gene expression level of our illustrative dataset support our proposal of a systems biological approach towards a molecular basis of heterosis.

3.2 Introduction

The phenomenon of heterosis has already been known since the last century (Shull, 1908). It was defined as "increased vigor, size, fruitfulness, speed of development, resistance to disease and to insect pests, or to climatic rigors of any kind, manifested by crossbred organisms compared with corresponding inbreds, as the specific results of unlikeness in the constitutions of the uniting parental gametes" by Shull (1952). This definition is restricted to describing the phenotypes that result when two different inbred lines are crossed. Therefore, it is often interpreted as not implying a genetic basis for heterosis (Lamkey and Edwards, 1999). This was accomplished by Schnell and Cockerham (1992) defining heterosis as the difference in performance between hybrid and the mean of the two parents. Figure 3.1 displays such a quantitative genetics definition of heterosis. Mid-parent heterosis is the difference in phenotype value between the heterozygous offspring and the mean of the homozygous parents, while best-parent heterosis describes the situation where the hybrid exceeds the best parent.

Three different genetic models to explain heterosis have been suggested: dominance (Bruce, 1910; Xiao et al., 1995), overdominance (Shull, 1908; East, 1936; Crow, 1952)

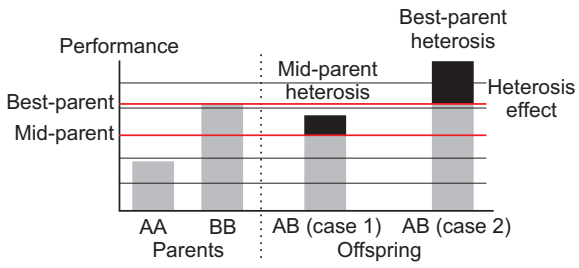


Figure 3.1: Quantitative genetics definition of heterosis. The black parts of the performance of the heterozygous offspring denote the heterosis effect

and epistasis (Schnell and Cockerham, 1992; Li et al., 2001; Luo et al., 2001). These hypotheses can be divided into approaches based on dominance or overdominance and global approaches (epistasis) (for review see Lamkey and Edwards (1999) and Birchler et al. (2003)). Towards a molecular basis of heterosis, it has been analyzed which genes show the above genetic non-additivity in their expression levels (Vuylsteke et al., 2005), and if such genes are enriched in yield-related QTL (Wei et al., 2009). However, a molecular mechanistic model, which would be able to explain how the observed phenomena on the molecular level are integrated to result in heterosis on the phenotype level, is still lacking.

In our contribution, we use a systems biology approach to analyze heterosis in *Arabidopsis thaliana* plants based on patterns in genome-wide gene expression profiles. Already Robertson and Reeve (1952) suggested that heterozygotes are likely to posses a greater biochemical versatility by carrying greater diversity of alleles. Additional alleles at heterozygous loci may lead to additional regulatory interactions in the molecular network. Equipped with an enlarged repertoire of regulatory possibilities, hybrids may possibly be able to correctly respond to a higher number of environmental challenges leading to higher adaptability (individual acclimation ability) and, thus, the heterosis phenomenon.

Nowadays, high-throughput techniques, such as microarrays, allow measuring genome-wide feature profiles simultaneously. In global approaches, these datasets can be used to discover the interactions of molecules, how they are organized in networks and how the different networks are linked to each other (Barabási and Oltvai, 2004). Partial correlations have been recommended to estimate regulatory interactions from observational data (Werhli et al., 2006).

Simple network models have been proposed to model the regulatory apparatus in a parsimonious way (Shubik, 1996; Somogyi and Sniegoski, 1996; Genoud and Métraux, 1999). On this background we developed our "network hypothesis of heterosis" (Andorf et al., 2009). Our conceptual modeling results proposed that higher adaptability comes with an increased number of molecular interactions. To characterize the global interaction structure of regulatory networks, we use partial correlations (association networks). Based on the hypothesis of Robertson and Reeve (1952) and our conceptual model, we expect that the heterozygous genotypes show enriched partial correlations compared to the homozygous parents. These larger partial correlations represent the additional regulatory interactions in the molecular networks of the hybrids. Also, a gene set enrichment analysis is included to check for pathway-specific enlarged partial correlations.

The hypothesis was already tested on a metabolite dataset of samples of *A. thaliana* plants (Andorf et al., 2009). In this paper we will check if the hypothesis also holds true for gene expression data of the same genotypes. We use a certainly limited dataset, but aim to propose and illustrate a systems biological view which allows for an integrated hypothesis about the molecular basis of heterosis, complementing single gene and quantitative genetics approaches.

3.3 Materials and methods

3.3.1 Experimental data and preprocessing

Gene expression data were measured using Agilent's *Arabidopsis thaliana* Microarray Kit 4x44k, P/N G2519F (Agilent Microarray Designs ID 021169, arrays contain four subarrays where each represents a different hybridization). To isolate the RNA the innuPREP Plant RNA Kit (845-KS-2060250, Analytik Jena) was used. The RNA was obtained from seedlings of *A. thaliana* of two homozygous lines C24 and Columbia (Col-0) and the reciprocal crosses C24 × Col-0 and Col-0 × C24. Gene expression profiles were measured during early development at seven time points [4, 6, 10, 15, 20, 25 and 30 days after sowing (DAS)]. For each measurement, a group of seedlings (Petri dish, pot) was grown and fully harvested after every specific time of growing.

Figure 3.2 shows the experimental design, a multiple nested loop design. Each arrow represents one subarray, where the arrowhead symbolizes that the sample was labeled with one color and the root of the arrow symbolizes the other color. For each genotype-time point combination 2 or 4 biological replicates were measured. For the time points of 4, 10, 20 and 30 DAS, we had four replicates each. Part of the subarray that contains the samples of C24 at the time points 15 and 20 DAS (dashed

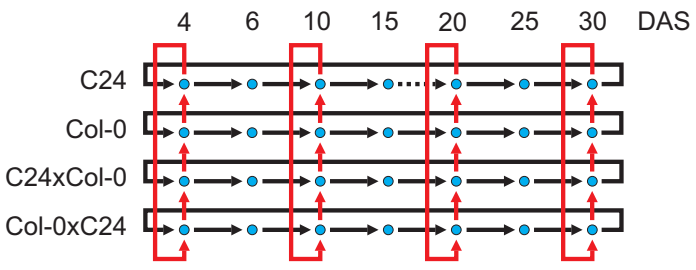


Figure 3.2: Experimental design with multiple loops. Each arrow symbolizes one subarray (*dashed arrow*: subarray was excluded from analysis)

Table 3.1: Eight Boolean variables related to outliers (Agilent Technologies Inc., 2008)

Green channel	Red channel
gIsFeatNonUnifOL	rIsFeatNonUnifOL
gIsBGNonUnifOL	rIsBGNonUnifOL
gIsFeatPopnOL	rIsFeatPopnOL
gIsBGPopnOL	rIsBGPopnOL

arrow in Figure 3.2) was covered by an air bubble and therefore, this subarray was excluded from the analysis.

Figure 3.3 summarizes the workflow of our analysis, beginning with the raw data from these microarray hybridizations.

During reading the raw data with the function *read.maimages* of the Bioconductor (Gentleman et al., 2004) *R* package *limma* (Smyth, 2005), low quality spots were detected using eight quality features (see Table 3.1) described in the reference guide of the Agilent Feature Extraction Software (Agilent Technologies Inc., 2008). Afterwards, the raw intensities of the spots that were not flagged out, were background corrected using the method *normexp* (Ritchie et al., 2007) of the *R* package *limma*. Background corrected values were lowess normalized to get as unbiased red/green-ratios as possible. For global comparability, the data of all arrays were quantile normalized (Smyth and Speed, 2003). For 3651 genes, more than 20% of the measured values were flagged out and therefore these genes were excluded from further analysis.

As proposed by Yang et al. (2002), normalized gene intensities were obtained as in Eqs. 3.1 and 3.2 from re-parameterizing the normalized log-ratios (M) and mean log-intensities (A) from the *limma* analysis.

$$M = \log I_{Cy5} - \log I_{Cy3} \qquad A = \frac{1}{2}(\log I_{Cy5} + \log I_{Cy3}) \tag{3.1}$$

$$\log I_{Cy5} = A + \frac{1}{2}M \qquad \log I_{Cy3} = A - \frac{1}{2}M. \tag{3.2}$$

Regarding the locus IDs, 6,647 genes are represented by two or more spots on each subarray. The normalized intensity values for all measurements of these genes are replaced by the average of the values of the multiple spots. This leaves 33,445 genes for the further analysis.

Subsequently, profiles of non-expressed genes as well as approximately constant pro-

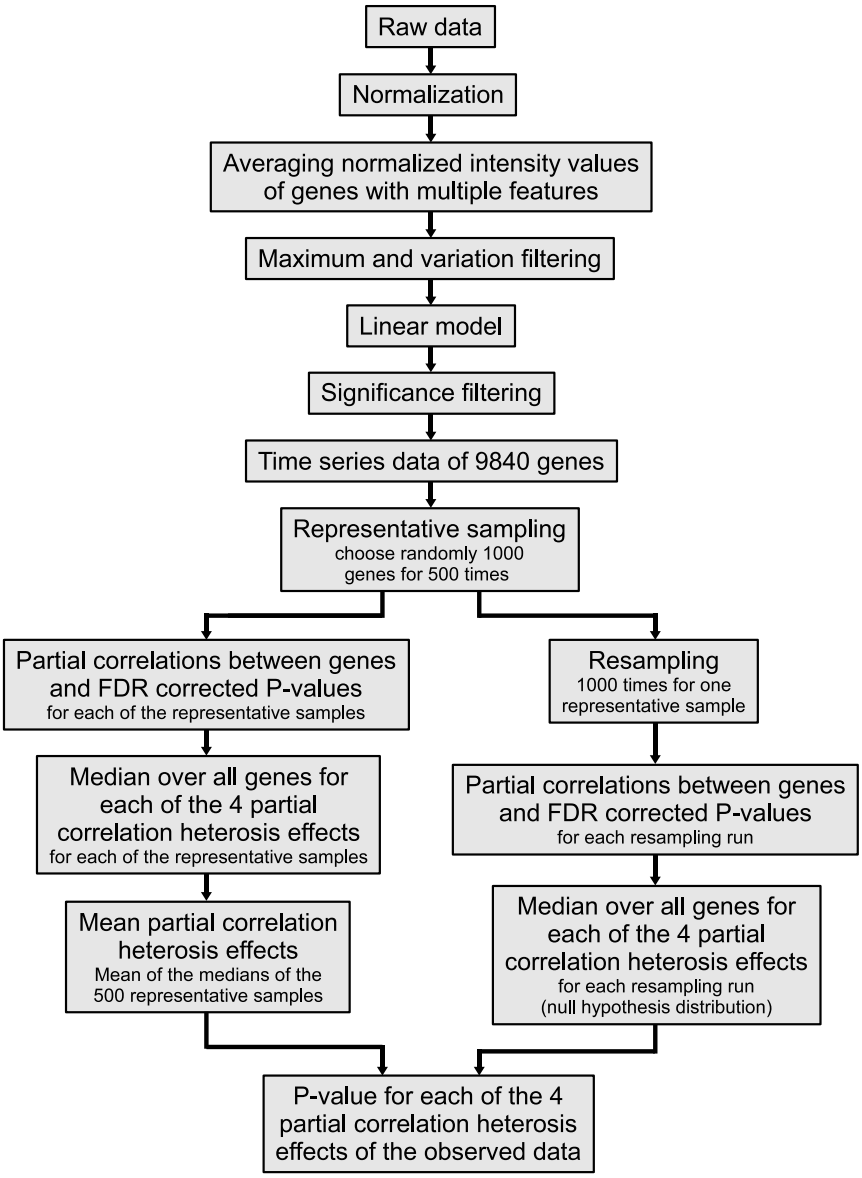


Figure 3.3: Workflow of our analysis

files were cleaned out applying two further filtering steps. In the first step, genes with very small maximum intensity values were screened out. Based on the distribution of the maximum intensity values of all genes (data not shown), we required a minimum intensity of $\log I \geq 7$ for at least one measurement of the gene. In the second step, genes were excluded from the subsequent analysis which showed low variation in their normalized intensities. In this filtering step we applied a cutoff of 2.8.

We used a linear model (adjusted to Kerr et al. (2000); Kerr and Churchill (2001)) to analyze the experimental loop design and estimate the gene expression profiles. It contains the factors g denoting the four genotypes, factor $t \in \{1, ..., 7\}$ denoting the seven time points of the developmental time series, their interaction $g \times t$, factor $AR \in \{1, ..., 11\}$ denoting the array (containing the four subarrays) and factor $DcRNA \in \{1, ..., 7\}$ denoting the date of cRNA synthesis. The fitting of the linear regression was done on a gene-wise basis for the following model where $y_{i,j,k,l,m}$ depicts the normalized gene intensities.

$$y_{i,j,k,l,m} = \mu + g_i + t_j + (g \times t)_{i,j} + AR_k + DcRNA_l + \varepsilon_{i,j,k,l,m}. \tag{3.3}$$

In this model, μ gives the overall gene-wise mean, the four genotypes are denoted with index i, the seven time points with index j, the array with index k, the date of cRNA synthesis with index l and the replicates with index m (between 1 and 4 biological replicates; see Figure 3.2 for details). A factor *dye* was not included in this model because it was not significant. Estimated gene expression values, $y^*_{i,j}$, were obtained from model 3.3 as in Eq. 3.4

$$y^*_{i,j} = g_i + t_j + (g \times t)_{i,j}. \tag{3.4}$$

Afterwards, we applied an additional filtering step on the estimated effects of the linear model. In this significance filter we filtered out genes that do not show a significant time and/or genotype-time interaction effect. We corrected the P-values for these effects using the FDR correction described by Benjamini and Hochberg (1995). We choose a liberal cutoff of 0.2 as significance level to only exclude genes which show nearly no time dependency or $g \times t$ interaction. After this filtering step, 9,840 genes remained for all further analyses, a number inline with our expectations from earlier expression studies in *A. thaliana* (Ma et al., 2005).

The analyses were performed using *R* (R Development Core Team, 2008) (version 2.8.1) on an openSUSE Linux 11.0 (x86_64) server with 32GB RAM. Raw gene expression data, estimated profiles as well as scripts are available upon request.

3.3.2 Network statistics

Werhli et al. (2006) suggested that partial correlations of features of time series profiles can be used to study causal regulatory interactions. Simulation results for metabolite time series data confirmed this (Andorf et al., 2009). So, we based our investigation of additional regulatory interactions in hybrids on the estimation of regulatory interactions through partial correlations. To calculate partial correlations we employed the approach as proposed by Opgen-Rhein and Strimmer (2007b). Their algorithm is implemented in the *R* package *GeneNet* (Opgen-Rhein et al., 2007). We used this package to obtain partial correlations from the normalized gene intensities of the seven time points. In *GeneNet*, partial correlations are calculated as in Eq. 3.5

$$\tilde{\rho}_{a,d} = \frac{-\omega_{a,d}}{\sqrt{\omega_{a,a}\omega_{d,d}}} \tag{3.5}$$

$\tilde{\rho}_{a,d}$ is the partial correlation between the genes a and d. $\omega_{a,d}$ is the element of the inverse covariance matrix. It is estimated using a shrinkage approach (Schäfer and Strimmer, 2005b) within the package *GeneNet*. For the shrinkage estimator of the partial correlations we used the default option "static" in the method *ggm.estimate.pcor* of the package *GeneNet*. Because we do not include any a priori information about the partial correlations in the shrinkage process, the covariance matrix is shrunk towards the identity matrix. To demonstrate the validity of this estimation procedure for the dimension of our data we conducted a methodology simulation study.

1. Construction of covariance matrices with constant covariance values of 0.25 and 0.4 for 1,000 nodes (the diagonal values were set to unity).

2. Cholesky decomposition approach (Parrish et al., 2009) to simulate gene expression data out of these matrices for seven time points.

3. Calculation of the partial correlations using the *R* package *GeneNet*.

4. Calculation of the difference between the mean of the partial correlations from the 0.4 covariance matrix and the one with 0.25 values.

5. Repeat of 1. - 4. for 100 times. The difference between the mean of the partial correlations of both simulated gene expression data had the same order of magnitude as the differences between the mean of partial correlations of the homozygous and heterozygous genotypes in our experimental data.

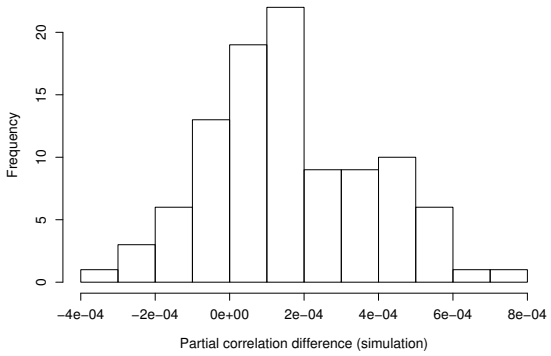


Figure 3.4: Differences between the mean of the partial correlations calculated from the simulated 0.4 covariance matrix and the one with 0.25 values for each of the 100 repeats. A positive difference was detected for 77% of the repeats

The simulation study described above is capable of showing that we are able to use the shrinkage estimator of the partial correlations as implemented in the package *GeneNet* in our study in a valid way. The resulting histogram of the differences between the mean partial correlations calculated from the 0.4 and the 0.25 covariance matrix for each of the 100 repeats is shown in Figure 3.4. In 77 cases from the 100 repeats, the mean difference of the simulated data was positive. In these cases, the stronger correlated data (0.4) lead to a detection of larger partial correlations in our simulation study. The means of all partial correlation values of the 100 repeats for the 0.25 and 0.4 covariance matrices, respectively, were 6.5×10^{-4} and 8.1×10^{-4}, respectively. For 1,000 randomly chosen genes of our experimental data, we calculated a mean of the means of the partial correlations for all four genotypes of 8.5×10^{-5}. Thus, we have shown that the shrinkage approach for the estimation of partial correlations by Schäfer and Strimmer (2005b) can be used for the dimension of 1,000 nodes and 7 time points as in our data. The power to identify enriched partial correlations in our simulation was 77%.

Within *GeneNet*, two-sided *P*-values for the test of non-zero correlation (null hypothesis: zero partial correlations) are calculated (Schäfer and Strimmer, 2005a; Strimmer, 2008). The *P*-values were corrected using the FDR correction described by Benjamini and Hochberg (1995). Like Werhli et al. (2006), we are interested in roughly estimating which regulatory interactions exist. Therefore, regarding our hypothesis that heterozygous genotypes contain more regulatory interactions, our focus is on the number of existing regulatory interactions, estimated as significant

partial correlations, and not on the value of each partial correlation itself. Hence, the further analysis of mid-parent and best-parent heterosis effects is based on s-values that are calculated like in Eq. 3.6

$$s_{b,f,u} = 1 - P_{b,f,u}^{FDR} \tag{3.6}$$

$P_{b,f,u}^{FDR}$ donates the FDR estimates according to Benjamini and Hochberg (1995) for the partial correlation between two genes ($f, u \in \{1, ..., N\}$, N genes in the analysis) of genotype $b \in \{\text{C24} \times \text{C24}, \text{Col-0} \times \text{Col-0}, \text{C24} \times \text{Col-0}, \text{Col-0} \times \text{C24}\}$. Using s-values, we get a high value for regulatory interactions that are most probably present (low corrected P-value) and low values for regulatory interactions that are probably not present in the regulatory network (high corrected P-value).

To determine the partial correlation mid-parent heterosis effect (see Figure 3.1) of each gene pair, we first calculated for each genotype separately for every single gene ($f \in \{1, ..., N\}$) the mean value of the s-values of its pairwise partial correlations to all other genes:

$$s_{mean,b,f} = \frac{1}{N-1} \sum_{u \in \{1,...,N\}, f \neq u} s_{b,f,u}. \tag{3.7}$$

Second, the mid-parent value for each gene was built out of the mean values calculated before for the homozygous genotypes:

$$s_{mid-parent,f} = \frac{1}{2} \sum_{v \in \{\text{C24} \times \text{C24}, \text{Col-0} \times \text{Col-0}\}} s_{mean,v,f} \tag{3.8}$$

Finally, the partial correlation mid-parent heterosis effects were calculated as the difference between the mean values from Eq. 3.7 of either hybrid and the mid-parent values:

$$\Delta s_{h,f,MPH} = s_{mean,h,f} - s_{mid-parent,f} \tag{3.9}$$

w denotes the respective heterozygous line ($h \in \{\text{C24} \times \text{Col-0}, \text{Col-0} \times \text{C24}\}$).

Simultaneously, we calculated the partial correlation best-parent heterosis effect values (see Figure 3.1). Here, instead of the mid-parent value, we determined the best-parent value (the maximum values of the mean values of the two homozygous genotypes; from Eq. 3.7):

$$s_{best-parent,f} = \max_{v \in \{\text{C24} \times \text{C24}, \text{Col-0} \times \text{Col-0}\}} s_{mean,v,f} \tag{3.10}$$

Afterwards, the partial correlation best-parent heterosis effect values were calculated as the difference between the mean values from Eq. 3.7 of either heterozygous genotype and the best-parent values. h denotes again the heterozygous line ($h \in \{\text{C24} \times \text{Col-0}, \text{Col-0} \times \text{C24}\}$):

$$\Delta s_{h,f,BPH} = s_{mean,h,f} - s_{best-parent,f} \tag{3.11}$$

As calculating partial correlations involves large matrices and, hence, a lot of working memory, we were not able to analyze partial correlations for all 9,840 genes that remain after filtering. Instead, we selected representative samples of 1,000 randomly chosen genes. This is displayed in the left chain of the workflow in Figure 3.3. To show that randomly selecting 1,000 genes indeed results in a representative sample, we selected 500 times randomly 1,000 genes and analyzed the variation of the features of interest. For each of the 500 repeats we calculated the partial correlation mid-parent and best-parent heterosis effects for either heterozygous genotype. For each of these four cases we determined the median of the calculated heterosis effect values. Thus, we got four median values for each of the 500 repeats; one median for each hybrid for the mid-parent as well as the best-parent heterosis effect. For each of the four cases, we then calculated the mean of the before determined 500 median values, $\overline{\Delta s_{\text{median},r}}$ (r indexing the four cases), as well as the 95% confidence interval (2.5 and 97.5% quantiles). If these confidence intervals exclude the value of zero partial correlation heterosis effects, we would be confident both to be able to show a robust effect and that our sampling approach yields representative samples in our sense.
To determine the significance of the observed partial correlation heterosis effects, we resampled the data of one randomly drawn representative sample of 1,000 genes in such a way that the genotype origins of the data are randomly re-assigned (right chain in Figure 3.3). For each gene in the set of 1,000, the estimated time profiles of the four genotypes were randomly re-assigned to the four genotypes (with replacement). This resampling was done 1,000 times. We calculated median values over the chosen 1,000 genes for each of the 1,000 resampling runs. This distribution of median partial correlation heterosis effects constitutes the null hypothesis distribution to establish a one-sided P-value for the originally observed partial correlation heterosis effects:

$$P^{\#} = \#(\Delta s_{\text{median,resampled}} \geq \overline{\Delta s_{\text{median},r}})/1,000. \tag{3.12}$$

A gene set enrichment analysis was performed to investigate if genes that show large partial correlation heterosis effect values (Eqs. 3.9, 3.11) are particularly enriched

in single pathways. We used gene sets (based on locus IDs) for 79 pathways. 30 of them were based on a MapMan annotation file (Usadel et al., 2009; Thimm et al., 2004), which, in turn, is based on the TAIR database version 8 (Swarbreck et al., 2008). 49 gene sets were built upon Plant Ontology (PO) terms (The Plant Ontology Consortium, 2002). Pathways that contained less than 10 or more than 4,000 of the genes we analyzed were excluded from this analysis, because too few genes in one pathway would make this pathway easily significant even if it just contains one or two genes with high partial correlation heterosis effect values. Too large pathways are not specific enough. The partial correlation heterosis effect values for each gene were determined using the first 100 representative samples of 1,000 randomly chosen genes each. Each time the mid-parent as well as best-parent heterosis effect values for either hybrid were saved and averaged. We got one partial correlation mid-parent and best-parent heterosis effect value per heterozygous genotype for each of our 9,840 genes. However, our gene set enrichment analysis was based on just 8,500 genes because for the other genes we did not have a locus ID and, thus, we could not assign them to the pathways. The median values of the mid-parent and best-parent heterosis effect values for either hybrid for all 8,500 genes are very close to the mean values shown in Figure 3.5.

We performed our gene set enrichment analysis using the hypergeometric distribution according to Drăghici et al. (2003) and Backes et al. (2007). This over-representation analysis measures enrichment by cross-classifying genes according to the membership in a functional category (gene set) and the membership in a selected list. We chose as selected list the 850 genes (10% of all genes in this analysis) that show the largest partial correlation mid-parent as well as best-parent heterosis effect for either heterozygous genotype. The resulting P-values were corrected using the FDR correction described by Benjamini and Hochberg (1995).

3.4 Results

As proposed by Werhli et al. (2006), an increase in molecular interactions can be measured as increase in partial correlations (also shown in a simulation study in Andorf et al. (2009)). Therefore, we investigated partial correlations according to Opgen-Rhein and Strimmer (2007b) of our experimental data, to test our hypothesis that regulatory networks of hybrids show enriched molecular interactions compared to their parental homozygous genotypes. The investigation was based on s-values (Eq. 3.6) to determine how many molecular interactions are probably present in the different genotypes.

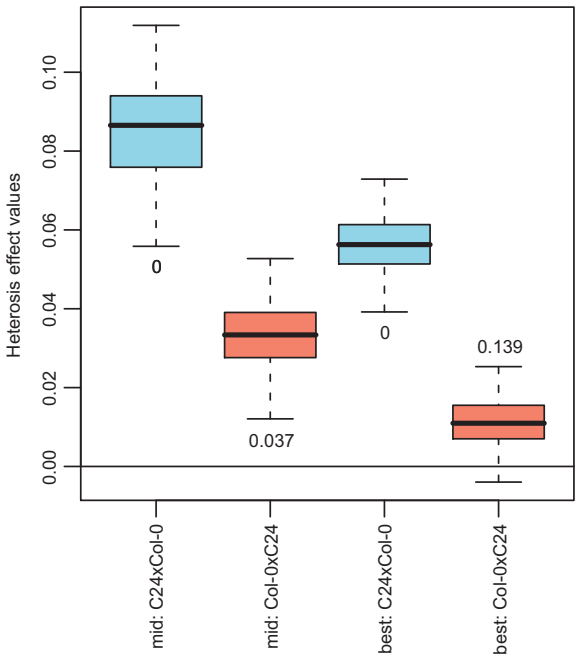


Figure 3.5: Distribution of the median values of the partial correlation heterosis effects of 500 repeated analysis of 1,000 randomly chosen genes. Mean values and 95% confidence intervals as well as the *P*-values are given

For 500 different sets of 1,000 randomly chosen genes we calculated the partial correlation mid-parent heterosis effect values (Eq. 3.9) as well as the partial correlation best-parent heterosis effect values (Eq. 3.11). Figure 3.5 displays the distribution of the 500 median values for the partial correlation heterosis effect values from the 500 repeated measurements of 1,000 genes (representative samples). Furthermore, the mean value and the 95% confidence intervall are shown for each case. For the heterozygous genotype C24 × Col-0, the 95% confidence intervals exclude the zero for the mid-parent as well as the best-parent partial correlation heterosis effect. The 95% confidence intervals for the genotype Col-0 × C24 exclude the zero just for the mid-parent partial correlation heterosis effect values and not for the best-parent partial correlation heterosis effect values. These three cases for which the 95% confidence intervals exclude the zero show the effect of enrichment of partial correlations in the transcriptome of the heterozygous lines and, furthermore, we are confident that choosing 1,000 genes randomly out of 9,840 genes leads to representative samples in this sense. For the last case we cannot decide on this basis if selecting 1,000 genes randomly is not representative or if this genotype does not show a best-parent partial correlation heterosis effect. We also determined the significance of the partial correlation heterosis effects of the observed data. This analysis was based on the re-

sampling of one set of 1,000 randomly chosen genes. For the genotype C24 × Col-0 we calculated a *P*-value of zero for the mid-parent as well as the best-parent partial correlation heterosis effect. Hence, both effects are significant for this heterozygous genotype. For the other heterozygous genotype (Col-0 × C24), only the mid-parent partial correlation heterosis effect is significant with a *P*-value of 0.037. For the best-parent partial correlation heterosis effect of this genotype, we determined a *P*-value of 0.139. Thus, the best-parent heterosis effect is not significant for the genotype Col-0 × C24. The *P*-values are also given in Figure 3.5.

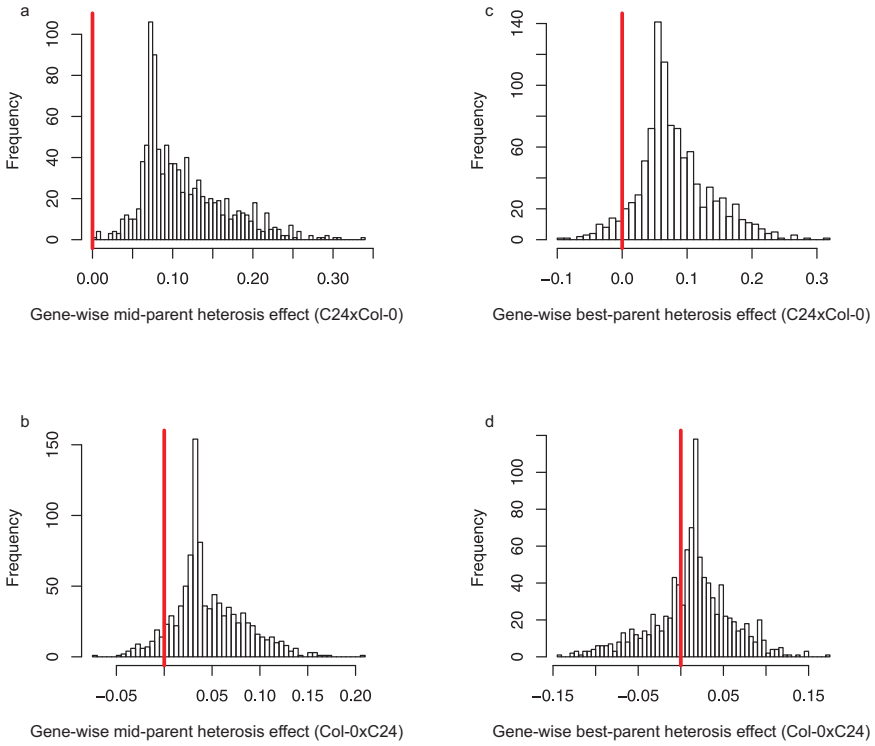


Figure 3.6: Display of partial correlation mid-parent heterosis effects (see Eq. 3.9) as well as partial correlation best-parent heterosis effects (see Eq. 3.11) for one representative set of 1,000 genes. For both hybrids most of the genes show a larger significance of the partial correlations than the mid-parent values or best-parent values, respectively

Figure 3.6 shows the partial correlation mid-parent as well as best-parent heterosis effect values for either heterozygous genotype for one set of 1,000 representative genes in detail. The histograms show that most of the partial correlation mid-parent heterosis effects for both heterozygous genotypes (C24 × Col-0: Figure 3.6a; Col-0 × C24: Figure 3.6b) are positive. The shift to the right is not as big for the partial correlation best-parent heterosis effects (C24 × Col-0: Figure 3.6c; Col-0 × C24: Figure 3.6d) as for the mid-parent heterosis effects but still noticeable.

Table 3.2: Results of gene set enrichment analysis for partial correlation heterosis effects: Enriched pathways of TAIR and PO

C24 × Col-0 mid-parent	Col-0 × C24 mid-parent	C24 × Col-0 best-parent	Col-0 × C24 best-parent
Male gametophyte	Male gametophyte	Male gametophyte	Male gametophyte
Stress	Lateral root primordium	Sperm cell	Lateral root primordium
Sperm cell	Transport	Stress	Sperm cell
Redox regulation	Sperm cell		
Photosynthesis	Primary root apical meristem		
	Lipid metabolism		
	Ovule		

In our gene set enrichment analysis we investigated if the partial correlation heterosis effects are enriched in some particular pathways. Table 3.2 shows the pathways that are enriched in either case of the partial correlation mid-parent and the best-parent heterosis effect for both hybrids.

3.5 Discussion

Our study aims at contributing to the understanding of heterosis at the molecular level by proposing a systems biological approach to analyze molecular profile data in hybrids. We estimated regulatory interactions between genes as partial correlations of their transcript profiles in a genome-wide approach for the early development of two homozygous *A. thaliana* lines and their reciprocal crosses. Results show a genome-wide global increase in the significance of partial correlations between transcript profiles in the hybrid lines as compared to the mid-parent as well as best-parent expectations. Moreover, in some functional groups of TAIR as well as PO terms both hybrid lines show a particularly high partial correlation heterosis effect. These results confirm earlier findings on the metabolite level (Andorf et al., 2009) and provide further support to a *molecular network hypothesis of heterosis* which we developed as systems biological approach contributing to a better understanding of the molecular basis of heterosis.

Within the existing diversity of explanatory hypotheses towards a molecular basis for heterosis, our approach aims to investigate changes in regulatory interaction on a

global level, rather than searching for single responsible loci. Existence of regulatory interactions on a global scale is estimated through significance of partial correlations. We therewith follow a line of argumentation taken as early as in the 1950s when Robertson and Reeve (1952) or Maynard Smith (1956) suggested that genetic heterozygosity might result in greater biochemical versatility in development and for reacting to environmental challenges. A larger repertoire in regulatory possibilities on the molecular level could result in the observed superior hybrid vigor. Also, recent discussions about possible molecular causes of heterosis include the notion of altered regulatory effects in hybrids and the positive effects of an enlarged repertoire of regulatory responses (Birchler et al., 2003; Song and Messing, 2003). The emphasis of our study is on substantiating this hypothesis as to enable to experimentally measure the enlarged regulatory versatility in hybrids as global structures on the molecular level.

In an earlier contribution, we proposed a systems biological approach contributing to an understanding of heterosis at the molecular level which we termed "network hypothesis of heterosis" (Andorf et al., 2009). Taking a very simplistic parsimonious view, we considered the Boolean network approach, following Genoud and Métraux (1999), to demonstrate how the enhanced possibility to correctly respond to environmental challenges is linked to an enlarged number of regulatory interactions. These were estimated as significant partial correlations of metabolite profiles in the same design as for the current study at some earlier time points of development. Summarizing the earlier results with those of the current study, we were now able to show a global enrichment of the number/significance of the partial correlations in the hybrid lines on both metabolome and transcriptome level for our illustrative datasets.

Regulatory interactions can only be estimated from correlation structures of a regulatory network in such parts where ongoing regulatory processes lead to measurable changes in the respective molecular profiles. As both datasets concern the early development of *A. thaliana*, where Meyer et al. (2004) showed that the foundations of biomass heterosis are laid, it might be speculated that it is the nature of this biomass phenotype that it concerns a global adaptation process of the seedling. Later developmental stages and adaptation processes, such as flowering, fruit ripening or other more specific phenotypes may require more local, limited molecular responses, e.g., restricted to special pathways or gene regulatory modules. The current study as well as the results of Andorf et al. (2009) mostly show *global* changes in partial correlation structures, i.e., increase in estimated regulatory interactions.

However, as result of our gene set enrichment analysis several gene sets appeared to be specifically enriched. We hypothesize that these genes are among the subset

of highly regulated genes during the specific developmental interval of our study. With the small-powered study design in mind, we do not want to speculate about biological interpretations of specific enriched gene sets.

In other species, such as *Drosophila* or mice, it became evident early in heterosis research that stress conditions were prone to cause pronounced heterosis effects (Harrison, 1962; Maynard Smith, 1956). A possible reason is that under such conditions the regulatory system is challenged to its limits. Hence, it is then necessary to make full use of the spectrum of regulatory possibilities. This may lead to inferior performance of the homozygous parental lines based on their limited regulatory possibilities when compared to their heterozygous offspring. In the setting of the current study, establishing a viable seedling under laboratory conditions, such as a climatic chamber opposed to the natural environment, may represent such an environmental challenge capable to show the enhanced potency of the hybrids' molecular regulatory repertoire.

When confronting hybrid genotypes with the environment, e.g., when recording performance in interesting environments for breeding and exploitation, *functional* data such as gene expression or metabolite profiles allow an additional, deeper characterization of potentially advantageous crosses. For example, Thiemann et al. (2010) search for gene expression signals of single genes correlated with hybrid performance in maize and functionally study their candidates using GO terms. Frisch et al. (2010) follow an alternative strategy. Parental gene expression values are used to build a distance measure which is used to predict hybrid performance with a linear model. Further approaches exist to combine functional and genetic data for hybrid performance prediction (Steinfath et al., 2010). Hence, these studies might complement respective results from QTL studies. As Melchinger et al. (2007b) found in their quantitative genetics study of *Arabidopsis* heterosis, QTL heterosis effects are to a large extent dependent on the whole genetic background. If a given genetic background is advantageous or not, certainly is dependent on the environment. This dependency is only accessible via *functional* tests. Several groups (Vuylsteke et al., 2005; Swanson-Wagner et al., 2006; Guo et al., 2006; Wei et al., 2009) investigated hybrids in comparison to their homozygous parents on the functional level, measuring genome-wide gene expression levels. In addition to their findings about proportions of realized modes of gene action in hybrids, our own contribution can be seen as proposing an idea for a systems biological heterosis analysis of the functional domain or gene expression level. We propose a hypothesis how molecular correlation structures specific for heterozygotes could be understood as mechanistic link between molecular and phenotypic manifestation of heterosis.

Considering regulatory interactions and possibilities to infer their global structure from molecular profile data, it is evident that a lot of existing regulatory interactions either involve molecular species which are not measured or act across different layers of the molecular regulatory apparatus profiled. Here, we adapt the view proposed by Somogyi and Sniegoski (1996), who emphasize the fact that the interactions deduced from molecular profiles of a specific level, e.g., metabolome or transcriptome, map regulatory processes of other molecular levels onto the one under consideration. Hence, the deduction of regulatory interactions for the specifically measured features may be wrong in detail, because the effects of molecules from other molecular levels are masked. This is especially so in the case of our study, as the number of time points sampled does not at all suffice to draw any strong conclusions on the level of a single estimated regulatory interaction. We think, however, that using our findings to build hypotheses about *global* structures of the molecular regulatory apparatus, such as an increased number of regulatory interactions in heterozygotes, is still allowed.

Partial correlations, also called *association networks*, are just one of several possibilities for estimating global regulatory interaction structures. The related so-called *relevance networks* (Butte et al., 2000), where Pearson correlations are measured to describe global correlation structures, are, however less eligible for our task. In contrast to partial correlations where indirect correlations are explicitly excluded, these remain an important factor when considering Pearson correlations. To emphasize this difference, it might be stated that when considering Pearson correlations it is save to talk about structures of *missing* correlations, whereas considering partial correlations reveals structures of *existing* correlations without being contaminated with indirect correlations. Werhli et al. (2006) recommended the use of partial correlations for the estimation of molecular interaction of regulatory networks from observational data, also contrasting it with a Bayesian network approach. In our study we follow this recommendation and use an algorithm proposed by Schäfer and Strimmer (2005b) which employs a shrinkage approach to estimate the partial correlations (*R* package *GeneNet*). Their approach is suitable for data with small sample size and large numbers of variables, as our genome-wide gene expression profiles. As we do not have any a priori information about the covariance structure of our transcriptome data, we chose the identity as canonical shrinkage target. Moreover, the time points of our time series data are not close enough in time to make additional use of the time series character – hence we chose the option "static" for application of the shrinkage estimator in the *GeneNet* package.

Time series data with only seven time points are a poor basis for investigating correlation structures of thousands of features. In our case we were concerned with

nearly 10,000 gene expression profiles from which we chose a set of 1,000 genes as representative sample. However, we refrained from interpreting partial correlations for single pairs of features, as due to the shortness of our time series, we were not able to carry out a more accurate network reconstruction analysis. Instead, we were interested in the global structures down to the level of a set of coarse grained pathways. This way, we feel that this kind of investigation of overall structure is still valid. A medium scale number of false positives or negatives may not disturb this coarse grained analysis results.

Regarding the number of features analyzed, it is necessary to also discuss the feature selection, or filtering process, which was performed previously to partial correlation analysis. Our filtering procedure has been chosen such as to filter out gene expression profiles which were likely representing features not expressed or regulated during the time interval of early development in our experiment. We chose cutoffs for the filtering process such that around 10,000 genes remained for further analyses. This number matches what is expected from existing studies regarding proportions of actively expressed genes in different tissues of *Arabidopsis* (Ma et al., 2005).

Our dataset is also a compromise from another point of view. The plant tissue used for feature extraction (RNA as well as metabolite isolation procedures) was the complete young seedling. Hence, we assessed only the average of tissues constituting the seedling. Inferences about the regulatory structure are therefore possibly exclusively valid on the global scale we address, most likely not for many specific single features and their correlations.

Furthermore, we are aware of the fact that only a single cross is a poor basis to draw general conclusions.

Summarizing the methodological considerations, it remains to emphasize that the dataset of the current investigation could be analyzed with valid results only at the coarse grained global level. However, at this level, gene expression as well as metabolite profiles jointly pointed towards an increase in the significance of partial correlations. This, based on Werhli et al. (2006), we interpret as increase in number of interactions allowing for an increased adaptability to environmental challenges during early seedling development.

Future investigations should certainly involve multiple lines, multiple species, multiple time windows of development or different environmental challenges for homozygous parents and their hybrids to be proven on the *functional* level. Also, longer time series should be investigated. Moreover, when more detailed time series data become available, an analysis of regulatory structures on a smaller scale could become possible where more valid investigations could be taken on the levels of special

pathways, regulatory modules or motifs (Hartwell et al., 1999; Milo et al., 2002; Lee et al., 2002). Also, integrative bioinformatic approaches involving the combination of gene expression with metabolite profiles and hQTL data could reveal promising results, especially for more local heterosis phenotypes affecting only a small part of the regulatory network (see for example Gärtner et al. (2009); Wei et al. (2009)). The discovery of functional groups of genes with particularly enriched partial correlations could complement and refine quantitative genetics analysis about non-additive gene actions and help to approach an understanding of the molecular basis of heterosis. Hence, the systems biological approach towards finding the molecular basis of heterosis introduced with the current investigation should be seen as a methodological proposal illustrated with a small dataset, complementary to the quantitative genetics approach, which is not taking into account global structures of the various OMICS levels, and the single-gene centered approaches, which involve data of much higher resolution for the price of neglecting the global view.

Acknowledgements

This work was supported by the German Research Council (DFG) under Grants RE 1654/2-1 and SE 611/3-1. We want to thank Dirk Hincha (MPIMP-Golm) and his lab for supporting our gene expression experiments.

4 Integration of a systems biological network analysis and QTL results for biomass heterosis in Arabidopsis thaliana

Sandra Andorf [1], Rhonda Christiane Meyer [2], Joachim Selbig [3], Thomas Altmann [2], Dirk Repsilber[1]

[1]: Research Unit Genetics and Biometry,
Leibniz Institute for Farm Animal Biology (FBN)
Wilhelm-Stahl Allee 2, D – 18196 Dummerstorf, Germany
[2]: Department of Molecular Genetics,
Leibniz Institute of Plant Genetics and Crop Plant Research (IPK)
Correnssstr. 3, D – 06466 Gatersleben, Germany
[3]: Institute for Biochemistry and Biology, University of Potsdam
Karl-Liebknecht-Str. 24-25, D – 14476 Potsdam-Golm, Germany

Under review

4.1 Summary

To contribute to a further insight into heterosis we applied an integrative analysis to a systems biological network approach and a quantitative genetics analysis towards biomass heterosis in early *Arabidopsis thaliana* development. The study was performed on the parental accessions C24 and Col-0 and the reciprocal crosses. In an over-representation analysis it was tested if the overlap between the resulting gene lists of the two approaches is significantly larger than expected by chance. Top ranked genes in the results list of the systems biological analysis were found to be significantly over-represented in the heterotic QTL candidate regions for either hybrid as well as regarding mid-parent and best-parent heterosis. This suggests that several genes that influence biomass heterosis are located within each heterotic QTL region.

Furthermore, the overlapping resulting genes of the two integrated approaches were detected to be particularly enriched in biomass related pathways. A chromosome-wise over-representation analysis gave rise to the hypothesis that chromosomes number 2 and 4 probably carry a majority of the genes involved in biomass heterosis in the early development of *Arabidopsis thaliana*. Our integrative approach allowed to identify candidate groups of genes, recognized by both approaches, which are likely to contribute to the molecular basis of biomass heterosis in early *Arabidopsis thaliana* development with enhanced specificity.

4.2 Introduction

The heterosis phenomenon, also known as hybrid vigor, was discovered in the early 20th century (Shull, 1908). It describes the superiority in fitness-related traits of F1 hybrids compared to their parental homozygous lines (Shull, 1948). Mid-parent heterosis (MPH) is the difference between the trait value of the hybrid and the average trait value of the two parental inbred lines, while best-parent heterosis (BPH) is the difference between the hybrid and the better of the homozygous parents. Even though the plant breeding interest in heterosis is high, the underlying genetic and molecular mechanisms are still not well understood.

In this work, we try to further approach the molecular basis of heterosis by integrating the results of our previously proposed systems biological hypothesis towards the understanding of heterosis on the molecular level (Andorf et al., 2009, 2010a) with the outcome of a quantitative genetics study for biomass heterosis in early development of *Arabidopsis thaliana* by Meyer et al. (2010). In both analyses the same two parental accessions, C24 and Columbia (Col-0), which are known to show biomass heterosis in their crosses (Meyer et al., 2004), were used.

Our proposed network hypothesis for heterosis (Andorf et al., 2009, 2010a) is based on partial correlations to characterize the global interaction structure of regulatory networks from observational time series data (Werhli et al., 2006). We expect a higher number of regulatory possibilities in the hybrids compared to the homozygous parents (Robertson and Reeve, 1952). This higher number of regulatory possibilities leads to more regulatory interactions in the heterozygous genotypes. According to our hypothesis, this increase in the connectivity of the regulatory networks can be detected as an increase in the significance of the partial correlations between the genes in either hybrid compared to the homozygotes. For each of the two hybrids we obtained a list of genes ranked according to the increase in significance of its partial correlation to each other gene compared to the mean of the parents (MPH) or the

better of the two parents (BPH).
The top ranked genes of each of these ranking lists are compared to the list of genes from QTL experiments which identified genomic regions involved in biomass heterosis (Meyer et al., 2010). Within the QTL regions, only a few genes causally related to biomass heterosis are expected. Therefore, we expected that the overlap between the genes identified in the QTL regions and the top ranked genes from the systems biological approach would *not* be significantly larger than by chance. Conversely, from the systems biological point of view, it is predicted that probably many genes are involved in the complex trait of biomass heterosis. Hence, if the two approaches towards finding genes responsible for biomass heterosis in the early development of *Arabidopsis thaliana* show a significantly larger overlap than by chance, it suggests that each of the identified heterotic QTL regions contains more than only a few genes influencing biomass heterosis. To test this, an over-representation analysis (ORA) based on the hypergeometric distribution was used in which the significance of the overlap between the resulting gene lists of either approach is calculated (Drăghici et al., 2003; Backes et al., 2007).
To analyze the distribution of genes contributing to biomass heterosis over all five *Arabidopsis thaliana* chromosomes, we ran a chromosome-wise ORA. Furthermore, ORA were applied to identify pathways which contain significantly more of the genes of the resulting candidate group of genes from both approaches than expected by chance.

4.3 Results

We performed an over-representation analysis (ORA) to analyze if two different approaches towards biomass heterosis in *Arabidopsis thaliana* point to similar genes which are probably responsible for this heterotic phenotype. A significant enrichment of the resulting genes from one analysis in the other would suggest that this assumption is true and, therefore, more genes influencing biomass heterosis are within the identified heterotic QTL regions than expected. Each of the analyses was performed for the two heterozygous genotypes C24 × Col-0 and Col-0 × C24 as well as regarding MPH and BPH.
Our ORA (setup shown in Figure 4.1) was based on a reference set of all $m = 33239$ *Arabidopsis thaliana* genes in the TAIR database version 9 (Huala et al., 2001). The test set was built out of the $n = 3133$ genes within the genomic regions that are involved in biomass heterosis determined in the quantitative genetics study by Meyer et al. (2010). Following Fury et al. (2006), we used different numbers of genes in

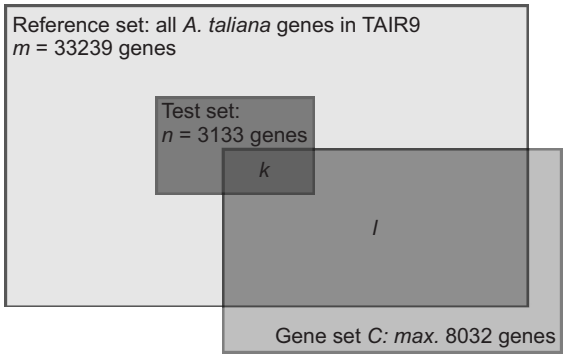


Figure 4.1: Setup of the over-representation analysis to test if a systems biological approach towards heterosis (Andorf et al., 2010a) and a quantitative genetics approach (Meyer et al., 2010) point to similar genomic regions influencing biomass heterosis in the early development of *Arabidopsis thaliana*.

the gene set. Each gene set was created from the genes with the x largest Δs-values (partial correlation heterosis effect values according to Eq. 4.2 and 4.3) separately for either hybrid as well as MPH and BPH. A large Δs-value indicates that the gene was identified in our systems biological analysis as probably involved in biomass heterosis in the early development of *Arabidopsis thaliana* (Andorf et al., 2010a). For each number x of genes (ranging from 0 to 8032 genes by steps of 100) we determined four gene sets $C_{x,h,a}$ ($h \in \{\text{C24} \times \text{Col-0}, \text{Col-0} \times \text{C24}\}$ and $a \in \{\text{MPH, BPH}\}$).

The ORA was used to estimate the probability that the number of genes which overlap between the test set and the respective gene set is either due to chance or represents a true enrichment of the genes of the gene set in the test set.

Figure 4.2a shows the results of the ORA. The x-axis depicts the number x of genes which were used in the gene set of each particular ORA. The y-axis shows the probability (P-value according to Eq. 4.5) of having as many or more than the observed $k_{x,h,a}$ genes in the overlap between the 3133 genes in the test set and the x genes in the respective gene set if the genes of the test set would have been chosen randomly out of the reference set. For genotype C24 $\times$ Col-0 for all gene sets $x \geq 2700$ significantly (significance level 0.05) more genes were detected in both approaches than expected by chance for MPH as well as for BPH (Figure 4.2a). P-values were only calculated for the case where the observed number of genes in the overlap between test set and gene set was larger than we expected just by chance.

The result for the other hybrid, Col-0 $\times$ C24, was similar. For $x \geq 1800$ for MPH and $x \geq 1600$ for BPH all ORA showed a significant overlap (significance level 0.1) between the results of the two different heterosis analyses. Different from genotype

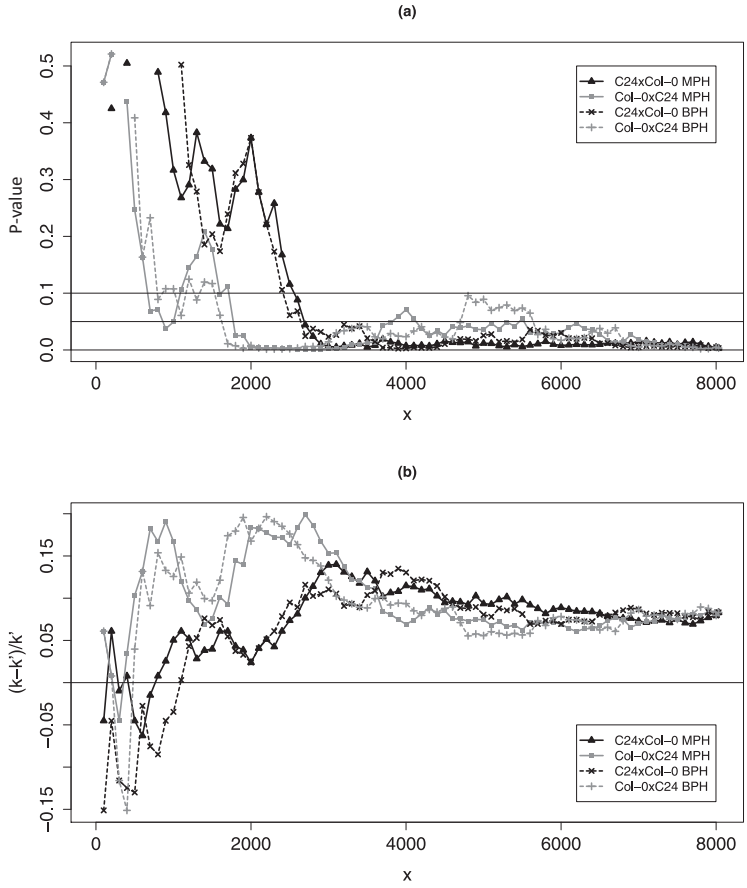


Figure 4.2: ORA results for gene lists of two different approaches towards biomass heterosis. x: number genes in gene set. (a) For gene sets of $x \geq 2600$ genes, both hybrids (C24 × Col-0: black; Col-0 × C24: gray) show a significantly larger overlap between test set (3133 genes determined in QTL mapping experiments for biomass heterosis (Meyer et al., 2010)) and gene set (determined in systems biological network analysis (Andorf et al., 2010a)) than expected for a random test set for MPH as well as BPH. (b) Fraction of how much more genes are observed in the overlap between test set and each of the gene sets than expected if this overlap would be a chance event relative to the expected overlap. The over-representation is significant (a) but not very strong (b).

C24 × Col-0, not all gene sets with more than a certain number of genes resulted in a P-value < 0.05. For MPH gene sets of $3800 < x < 4200$ genes and for BPH gene sets of $4700 < x < 5700$ led to P-values between 0.05 and 0.1 (Figure 4.2a).

Figure 4.2b shows the fraction of how much more genes were observed in the overlap between test set and gene set (k genes) than expected just by chance (k' genes) in relation to k'. For large x-values we determined a constant percentage of more genes than expected by chance.

Summarizing, we identified a significant over-representation of the gene set in the test set (Figure 4.2a), but the enrichment was not very strong. A maximum of around 20% more genes in the overlap than expected just by chance was detected and an average of a little less than 10% for gene sets of 4000 genes or more (Figure 4.2b).

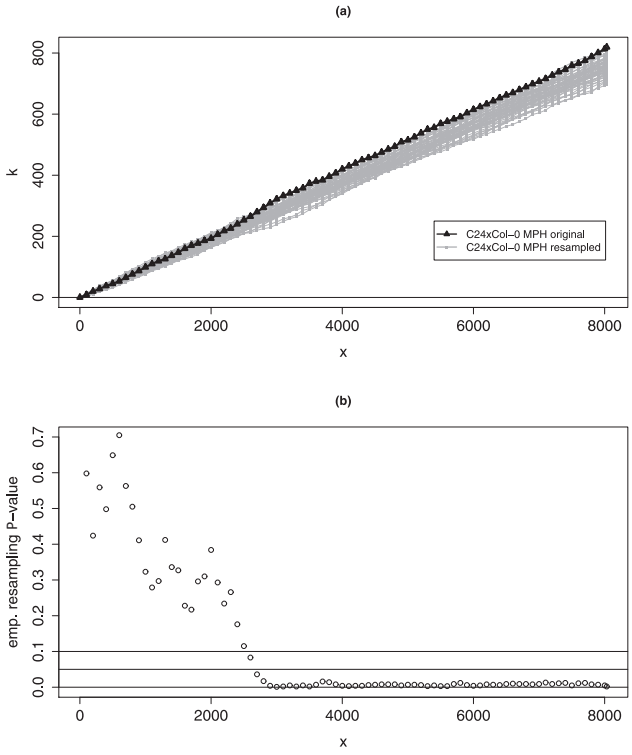


Figure 4.3: ORA with original and several random gene sets to confirm a significant over-representation between genes of two different approaches. (a) shows for C24 × Col-0 MPH the number (k) of genes in the overlap between 3133 genes in the test set and x genes in the "original" gene sets (black) and in 50 randomly chosen gene sets (gray), respectively. For $x \geq 2700$ the calculated empirical resampling P-values on the basis of 1000 randomly chosen gene sets confirm a significant over-representation of the gene set in the test set (b).

As a second validation of the significance of the observed enrichment, we ran a resampling analysis in which the genes in the gene set were sampled 1000 times randomly out of all genes in the reference set. Figure 4.3a shows the relation between the number of genes observed in the overlap between test set and each "original" gene set and in the overlap to the resampled gene sets for MPH of genotype C24 × Col-0. For the reason of clearness the results of only the 50 first resamplings were plotted. For $x \geq 2600$ the number of genes in the respective "original" gene set also present in the test set exceeded the overlap between test set and nearly each

randomly resampled gene set, confirming the findings from Figure 4.2a. This result is hardened by significant (significance level 0.05) corresponding empirical P-values, calculated for 1000 resamplings, for all $x \geq 2700$ (Figure 4.3b).

The calculation of the empirical resampling P-values for BPH C24 × Col-0 as well as MPH and BPH of Col-0 × C24 also confirmed the result of a significant over-representation for sufficiently large x shown in Figure 4.2a.

To study if the determined significant over-representation for sufficiently large gene sets is the same over the five chromosomes in *Arabidopsis thaliana*, we ran the ORA separately for each chromosome. Again, the genes from within the regions that were determined during the QTL analysis by Meyer et al. (2010) applying LOD-score thresholds for the empirical significance level of 5% were chosen as test set. As before, the gene sets were based on the results from the systems biological analysis by Andorf et al. (2010a). However, in this analysis each gene set $C_{x,h,a}$ was split up into five gene sets ($C_{x,h,a,chr}$). Each gene set contained only genes belonging to one of the five chromosomes.

The results of these chromosome-wise ORA are shown in Figure 4.4. For chromosomes number 1 (Figures 4.4a and 4.4b) and 5 no significant over-representation of the gene set in the test set was detected for either hybrid as well as heterosis measure, independent of the gene set size. For chromosome number 5 the observed overlap between the results of the two different approaches was for each number x of genes in the gene set smaller than expected just by chance. Therefore, no plots for chromosome number 5 are presented in Figure 4.4.

Chromosomes 2 (Figures 4.4c and 4.4d) and 4 (Figures 4.4g and 4.4h) showed a significantly larger overlap than expected by chance between test set and gene set for both heterozygous genotypes and both heterosis measures for nearly each gene set size. For chromosome 3 the result was not as clear as for the other ones. The hybrid C24 × Col-0 showed a significant enrichment (significance level 0.1) of the gene set in the test set for gene sets of 400 or more genes for MPH and BPH. The determined P-values for the ORA of the genotype Col-0 × C24 fluctuated with different gene set sizes between significant on the level of 0.05, significant on the level of 0.1 and not significant at all. Hence, for chromosome number 3 only C24 × Col-0 showed a significantly larger overlap between the results of the two approaches than expected by chance. However, this over-representation is much weaker than for chromosomes 2 and 4 (Figures 4.4d and 4.4h).

The plots in Figure 4.5 are based on the 3000 genes with the highest Δs-values for C24 × Col-0 MPH. The x-axis shows the genetic distance (Kosambi centimorgan (cM)) of each of the five *Arabidopsis thaliana* chromosomes (5a-e: chromosomes 1-5).

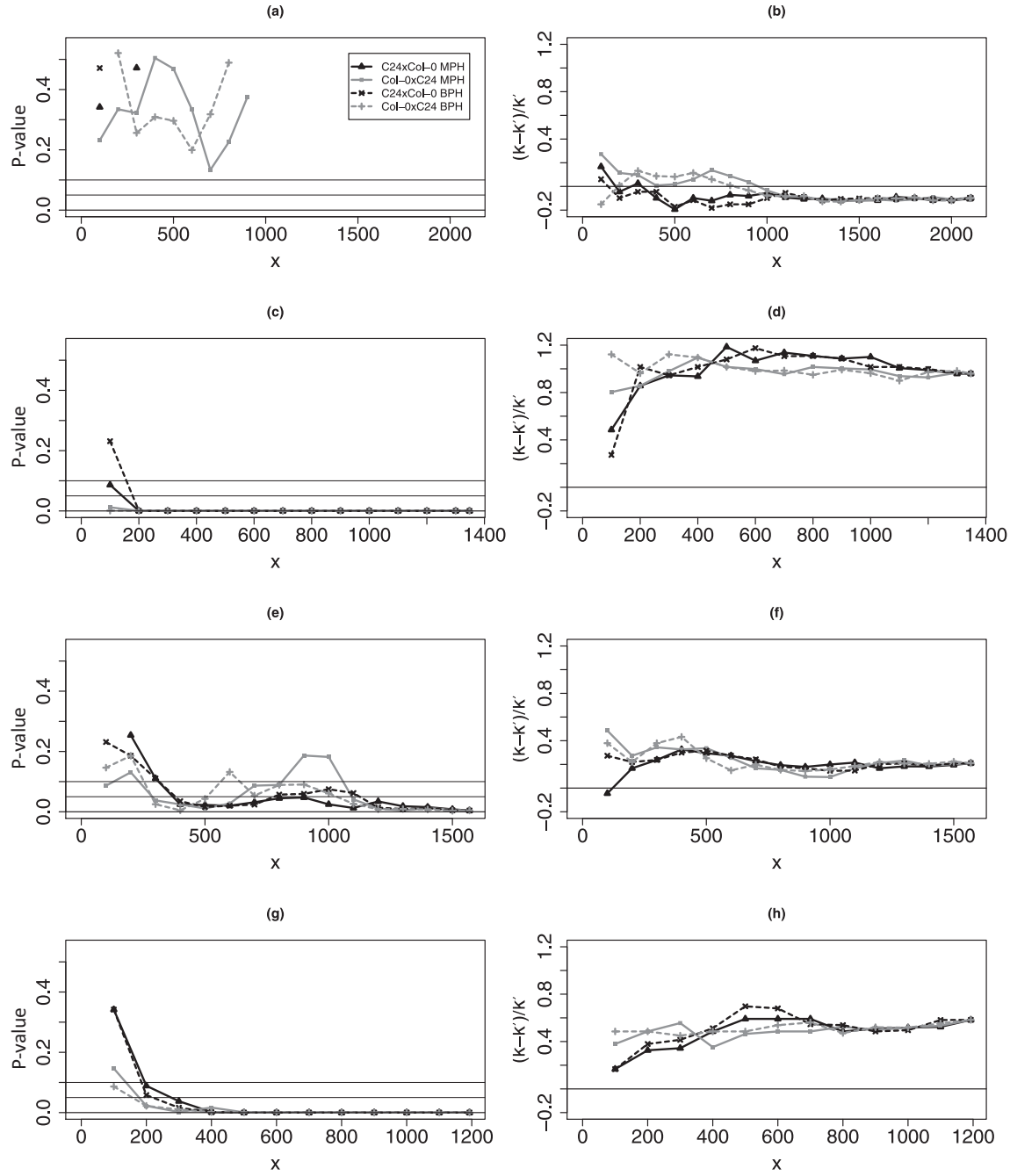


Figure 4.4: ORA results for gene sets that contain genes of only one of the five *Arabidopsis thaliana* chromosomes (x: number genes in gene set). (a)+(b): chromosome 1; (c)+(d): chromosome 2; (e)+(f): chromosome 3; (g)+(h): chromosome 4; for chromosome 5 no over-representation at all was observed and, therefore, no plots are shown. The probabilities of having as many or more genes in the overlap of random test sets to the gene set than observed for the experimental data are shown on the left side. The proportion of genes that were more in the overlap than expected by chance are shown on the right side. Genes on chromosomes 2 and 4 show a significant over-representation between the results of the two different approaches towards biomass heterosis in early development.

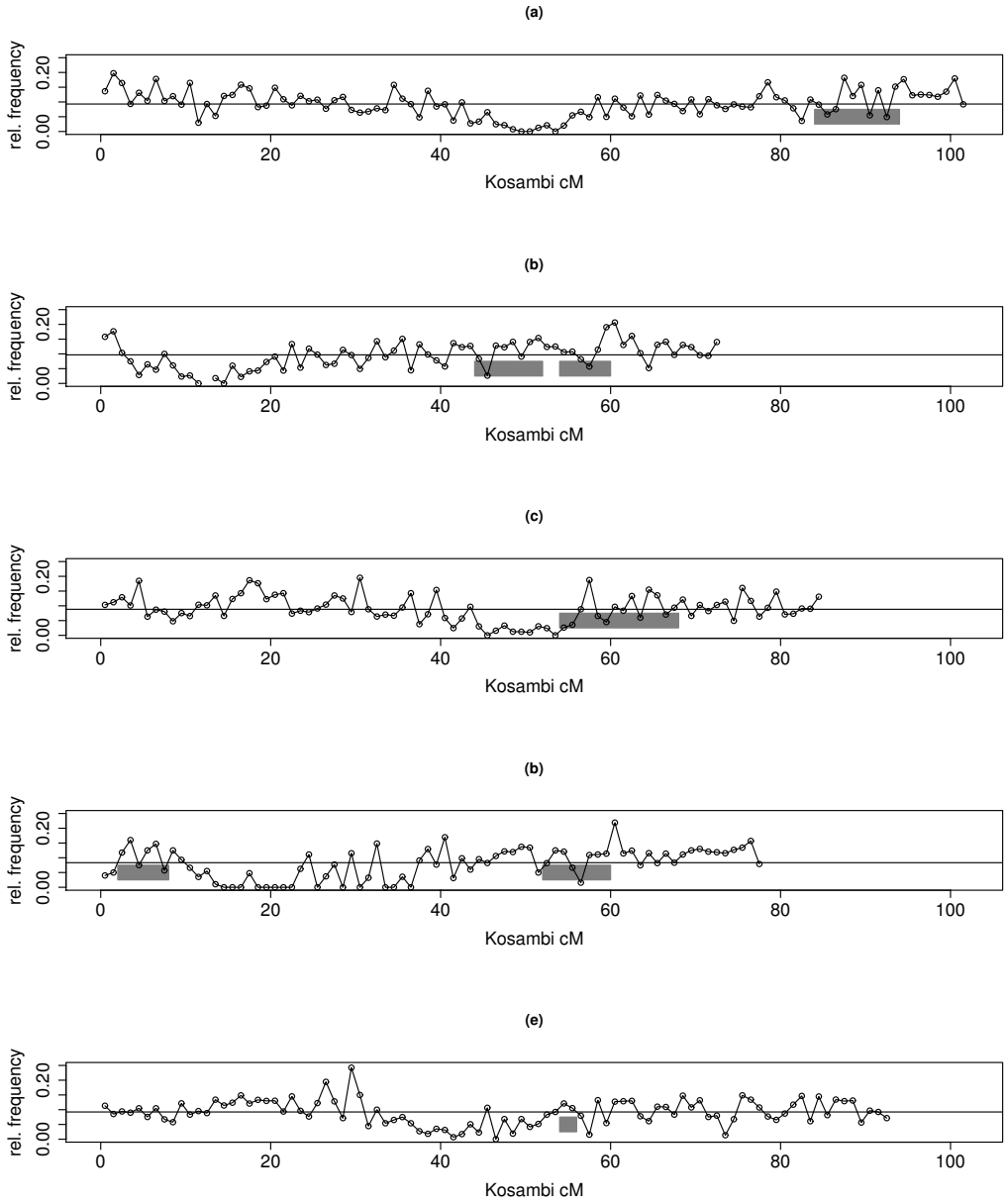


Figure 4.5: Relative frequencies (fraction between the number of detected 3000 top ranked genes of C24 × Col-0 MPH of the systems biological analysis and all known genes in *Arabidopsis thaliana* at a specific section on each chromosome) are plotted against the genetic distance (Kosambi cM). (a-e): chromosome 1-5. The gray boxes are the QTL candidate genomic regions contributing to biomass heterosis (Meyer et al., 2010). The horizontal lines are the median values of the relative frequencies per chromosome. For chromosomes 2 and 4 the relative frequency exceeds the median relative frequency especially in the area of the detected QTL. This is in line with the detected significant over-representation on these chromosomes of the genes detected in the systems biology approach in the genes from the QTL study.

The determined heterotic QTL regions are represented as gray boxes. To show that for some chromosomal sections more of the 3000 genes were detected than expected by chance, we calculated relative frequencies as the number of the 3000 genes in a certain section of Kosambi cM, divided by the number of all known *Arabidopsis thaliana* genes from TAIR9 in this section. By building the relative frequencies we accounted for the different gene densities at diverse chromosome regions. These relative frequencies are shown on the y-axis. The horizontal lines are the median values of the relative frequencies for each chromosome. The detected overall over-representation was weak (Figure 4.2b) but particular significant on chromosomes 2 (Figure 4.4c) and 4 (Figure 4.4g). In the plots of these two chromosomes (Figures 4.5b and 4.5d) more points are above the respective median relative frequency values in the areas of the QTL than in other regions on the respective chromosome.

Table 4.1: Ranked results of ORA of overlapping genes between two approaches towards biomass heterosis and pathways of TAIR8 and PO

C24 × Col-0 MPH	C24 × Col-0 BPH	Col-0 × C24 MPH
LP.08 eight leaves visible	LP.08 eight leaves visible	petiole
LP.12 twelve leaves visible	leaf lamina base	LP.02 two leaves visible
petiole	LP.12 twelve leaves visible	leaf lamina base
leaf lamina base	petiole	LP.12 twelve leaves visible
shoot	shoot	F mature embryo stage
LP.02 two leaves visible	LP.02 two leaves visible	LP.08 eight leaves visible
cotyledon	cotyledon	
male gametophyte	F mature embryo stage	
F mature embryo stage		
guard cell		

Furthermore, it was analyzed to which functional group within the *Arabidopsis thaliana* plants the genes in the overlap between the two approaches towards biomass heterosis in *Arabidopsis thaliana* belong. This was done by using an ORA with a different setup than before. In this analysis, the overlapping genes between the 3000 genes with the highest Δs-values, again separately for each hybrid and heterosis measure, and the 3133 genes identified in the quantitative genetics analysis were used as test sets. The reference set were all 8032 genes that were analyzed in the systems biological study according to Andorf et al. (2010a). Each of the 80 gene sets that were used contained genes that belong to one pathway based on PO terms (The Plant Ontology Consortium, 2002) or MapMan (Usadel et al., 2009) which in

turn is based on the TAIR8 database. Table 4.1 shows ranked lists of pathways with a significant over-representation (FDR corrected P-value < 0.05). For the test set of Col-0 $\times$ C24 BPH, no pathway with a significant enrichment was detected. The majority of pathways that showed a statistically significant over-representation are leaf or otherwise biomass related.
The fundamental data (such as the Δs-values) and results of the analyses are available upon request.

4.4 Discussion

In this study two different approaches to determine genes that contribute to biomass heterosis in early development of *Arabidopsis thaliana* were integrated. We could show that the quantitative genetics approach by Meyer et al. (2010) and the systems biological analysis by Andorf et al. (2010a) point to similar genomic regions influencing heterosis for biomass. An over-representation analysis (ORA) revealed that the resulting genes of these two studies showed a significantly larger overlap than expected by chance (Figure 4.2a). This result of a significant over-representation achieved in the parametric ORA was confirmed in a resampling analysis in which the genes of the gene set were randomly chosen 1000 times out of the reference set (Figure 4.3). However, while the enrichment was significant, it was not very strong (Figures 4.2b and 4.5). This result was achieved for either hybrid (C24 $\times$ Col-0 and Col-0 $\times$ C24) regarding both heterosis measures (MPH and BPH) (Figure 4.2) and markedly for two out of the five chromosomes (Figure 4.4).
In further ORA with a different setup we analyzed if the genes in the overlap between the results of the two approaches show a significant enrichment in one or more of 80 *Arabidopsis thaliana* pathways. The majority of pathways that showed a significantly larger overlap than by chance are leaf or otherwise biomass related (Table 4.1). This result is in line with an earlier analysis by Meyer et al. (2004), in which they detected heterosis in the trait of biomass in early development of the same *Arabidopsis thaliana* accessions that were under study in this work.
Some methodological details of our approach remain to be discussed. The number of genes determined in the quantitative genetics approach was fixed due to the preselected criterion of an empirical significance level of 5% for the LOD-score thresholds. The number of genes in the results list of the systems biological approach was not specified for the ORA. The ORA was run several times to study the overlap between the 3133 genes of the QTL mapping experiments (test set) and the top ranked 0 to 8032 genes of the systems biological analysis (gene set).

We did not run the ORA for a fixed size of the gene set because Fury et al. (2006) stated that the overlapping probability determined in ORA studies depends on the number of genes in the gene lists (test set and gene set) which are compared. Fury et al. (2006) have shown that the overlapping significance increases (P-values decrease) with increasing number of genes in the gene set (or test set). So, small gene set sizes have a small overlapping significance. This may be one of the reasons why the P-values for low numbers x of genes in the gene set are not significant in our analysis (Figure 4.2a). In contrast, if no true signal (over-representation) is present, the gene list size does not effect the P-values that are observed (Fury et al., 2006).
If Δs-values for each gene in the reference set (all *Arabidopsis thaliana* genes in the TAIR9 database) of our ORA were available, the analysis could have been extended to $x > 8032$ genes in the gene set. In this case we expect that the P-values of the over-representation would increase for larger gene set sizes x until it is not significant at all (Fury et al., 2006).
In the microarray experiments of the systems biological analysis 44k gene models were measured but reduced due to filtering steps and methodological reasons to only 8032 genes for this current work. The filtering involved a step in which genes that show no significant time and/or genotype-time point-interaction effect in the applied linear model were excluded from the subsequent analysis (Andorf et al., 2010a). This leads to the fact that in the 8032 genes in this study, the genes that are probably involved in biomass heterosis are already slightly enriched. This may be one reason for the significant over-representation for large x-values shown in Figure 4.2a.
Furthermore, we want to point out that the results of the chromosome-wise analyses have to be used with care because each of the gene sets was relatively small. This chromosome-wise analysis can, therefore, not give a firm insight into which of the chromosomes contain the genes that are mainly responsible for biomass heterosis.
In this work, an integrative analysis was presented. In comparison to the results of one experimental technique, the integration of two different experimental techniques accounts for the technological bias of each approach and the restriction to the particular level of biological information that is addressed by the single technique. The results that are found in an integrative analysis are more likely to be significant than the results of one single experimental technique as discussed in Steinfath et al. (2007).
However, no significant over-representation between the results of the two approaches integrated in this work could be expected because only a few genes per heterotic QTL region were assumed to influence biomass heterosis in *Arabidopsis thaliana*. Against this expectation, we could determine a significantly larger overlap between

the resulting candidate gene lists for biomass heterosis in the early development of *Arabidopsis thaliana* of the systems biological approach and the quantitative genetics analysis than expected by chance. Furthermore, the genes within this overlap are, in turn, significantly enriched in biomass related *Arabidopsis thaliana* pathways. This suggests that more genes (not only the expected few genes) from within each QTL region are somehow involved in biomass heterosis. So, most probably several genes in the respective regions led to the detection of each heterotic QTL region. Furthermore, this significant enrichment is in line with the hypothesis that functionally related genes are rather adjacent on the chromosomes than randomly distributed. Riley et al. (2007) proposed that the distribution of molecular functional classes of genes in *Arabidopsis thaliana* is not locationally independent. If functionally related genes influencing biomass heterosis would be distributed randomly over the chromosomes, no significant over-representation would have been detected in this work.

For biomass heterosis in *Arabidopsis thaliana* at 15 days after sowing the found QTLs account for only up to around 30% of the phenotypic variation (Meyer et al., 2010). This may be one reason, why the identified over-representation was weak even though it was significant. Perhaps, more genes influencing the phenotypic variation were identified using the systems biological approach but these genes were not detected in the QTL mapping experiments. Another reason for this weak over-representation may be that *both* analyses that are integrated in this work identified several genes that do not affect the phenotypic variation. These genes which are not directly involved in biomass heterosis most probably differ in both approaches and, therefore, they do not overlap in the integrative analysis. The exclusion of these not overlapping genes is one aim of integrative analyses and may on the other hand be the reason for the weakness of the detected over-representation in this study.

Hence, the result of the integrative analysis not only points to more genes within the heterotic QTL regions influencing biomass heterosis than expected but it also suggests that the identified overlapping genes can be seen, with an increased confidence compared to the results of only one experimental technique, as a candidate group of genes which are likely to be involved in the molecular basis of biomass heterosis of early development of *Arabidopsis thaliana*.

4.5 Experimental procedures

4.5.1 Genes according to a systems biological analysis

To contribute to the understanding of heterosis, we proposed a systems biological hypothesis on the basis of molecular network structures (Andorf et al., 2009, 2010a). Following Robertson and Reeve (1952) and Werhli et al. (2006), we expect in our network hypothesis for heterosis that heterozygous genotypes which show heterosis contain more regulatory interactions and, therefore, denser regulatory networks than the homozygous parents. These additional regulatory interactions lead to an increase in the significance of partial correlations of features in the regulatory networks of the hybrids compared to the homozygous genotypes (Andorf et al., 2010a).

This network hypothesis for heterosis was tested and confirmed on transcriptome profiles from seven time points during the early development of two different homozygous *Arabidopsis thaliana* accessions (C24 and Col-0) and the two corresponding hybrids (Andorf et al., 2010a). These heterozygous plants are known to show a heterosis effect in their biomass phenotype (Meyer et al., 2004).

For details about the experimental data and raw data preparation, see Andorf et al. (2010a). Slightly different from the analysis described in Andorf et al. (2010a), another ANOVA model was applied in this study along with a cutoff of 0.3 for the corrected P-values of the effects in the used linear model. This way only genes that show nearly no time dependency and/or genotype-time point-interaction were excluded.

Different from the analysis described in Andorf et al. (2010a), in which several representative samples of 1000 genes each were analyzed, all genes remaining after filtering were processed at once in this work. Separately for each genotype, the partial correlations of the time profiles for each pair of these genes were calculated using the R package *GeneNet* (Opgen-Rhein and Strimmer, 2007b; Schäfer et al., 2009). Along with the partial correlation values itself, two-sided P-values (null hypothesis: zero partial correlation) were calculated and FDR corrected according to Benjamini and Hochberg (1995).

Following Werhli et al. (2006), a significant partial correlation symbolizes a probably present regulatory interaction between the two belonging genes. In order to have a high value for two genes, between which most probably a regulatory interaction exists, we built s-values:

$$s_{b,f,u} = 1 - P_{b,f,u}^{FDR} \tag{4.1}$$

$b \in \{\text{C24} \times \text{C24}, \text{Col-0} \times \text{Col-0}, \text{C24} \times \text{Col-0}, \text{Col-0} \times \text{C24}\}$ denotes the genotype. $P^{FDR}_{b,f,u}$ is the FDR corrected P-value of the partial correlation between the two genes $f, u \in \{1, ..., 9263\}$.

Separately for each genotype for every gene the mean of its s-values to each other gene was computed. This way, we received one s_{mean}-value for every gene for each of the four genotypes. On the basis of these s_{mean}-values we calculated the partial correlation MPH values $\Delta s_{h,f,MPH}$ in respect of our network hypothesis as the difference between the s_{mean}-value of either hybrid to the mean of the s_{mean}-values of the homozygous genotypes:

$$\Delta s_{h,f,MPH} = s_{mean,h,f} - \frac{s_{mean,\text{C24}\times\text{C24},f} + s_{mean,\text{Col-0}\times\text{Col-0},f}}{2} \tag{4.2}$$

where $h \in \{\text{C24} \times \text{Col-0}, \text{Col-0} \times \text{C24}\}$ denotes the hybrid and f the gene.

Partial correlation BPH values ($\Delta s_{h,f,BPH}$) were calculated as the difference between the $s_{mean,h,f}$-value and the larger of the $s_{mean,g,f}$-values of the two homozygous genotypes ($g \in \{\text{C24} \times \text{C24}, \text{Col-0} \times \text{Col-0}\}$) for each gene f:

$$\Delta s_{h,f,BPH} = s_{mean,h,f} - \max_{g \in \{\text{C24}\times\text{C24},\text{Col-0}\times\text{Col-0}\}} s_{mean,g,f} \tag{4.3}$$

A high $\Delta s_{h,f,MPH}$-value is achieved when the gene is probably involved in more interactions in the regulatory network of the respective hybrid than it is expected in the mean of the two homozygous parents. Correspondingly, a high $\Delta s_{h,f,BPH}$ suggests that the gene is involved in more regulatory interactions in the hybrid than in the better of the two parents. Following our network hypothesis for heterosis, genes with high Δs-values are likely to be involved in biomass heterosis.

With the cutoffs used in this analysis, our hypothesis about additional regulatory interactions holds true for either hybrid regarding MPH as well as BPH. However, we want to state that this is not a direct candidate gene approach in the way that we expect that for example the 100 genes with the highest Δs-values are all involved in biomass heterosis. Instead of that, we can just establish that a list of genes with high Δs-values contains various genes that may have an impact on biomass heterosis.

In our over-representation analysis (ORA) only genes which are identified with an AGI code can be used. Therefore, genes with unknown AGI code were excluded from the enrichment analysis. Δs-values of different gene models of one gene according to the TAIR9 database were averaged.

This left 8032 genes with a known AGI code and a Δs-value for each hybrid and heterosis measure for the ORA.

4.5.2 Genes according to a QTL analysis

The genomic regions involved in early stage biomass heterosis were identified as described in Meyer et al. (2010), using composite interval mapping (CIM) as implemented in the QTL mapping software PLABQTL (Utz and Melchinger, 1996). Data were obtained from recombinant inbred line (RIL) populations (Törjék et al., 2006) derived from the same *Arabidopsis thaliana* accessions C24 and Col-0 as used in the systems biological analysis described above. For the analysis, 838 testcrosses between the homozygous parents and 429 RILs were used.
Genomic regions were identified as having an influence on biomass heterosis, if the corresponding LOD-score exceeded the LOD threshold with an empirical significance level of 5%. LOD thresholds were determined separately for each trait by 5000 permutations (Churchill and Doerge, 1994). The genes within the genomic regions were identified on the basis of the TAIR9 genome release (The Arabidopsis Information Resource, ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR9_genome_release, June 2010) (Huala et al., 2001). 3133 genes in 7 genomic regions constituted the test set for our ORA.

4.5.3 Over-representation analysis

In an over-representation analysis (ORA), it is studied if a list of genes (gene set) is over-represented (represented more than expected by chance) or under-represented (represented less than expected by chance) with respect to another gene list (test set). Furthermore, the probability is estimated how likely this over-representation or under-representation is due to chance considering a specific reference set of genes (Drăghici et al., 2003; Backes et al., 2007). In this work the experimental data is tested for over-representation only.
We used an ORA to test if two different approaches towards more insight into biomass heterosis in *Arabidopsis thaliana* point to similar genes. One approach was based on quantitative genetics (Meyer et al., 2010) and the other on systems biology (Andorf et al., 2010a). Each of these two studies led to a list of genes probably involved in biomass heterosis. While the QTL approach provided directly a list with 3133 genes due to the given LOD-score thresholds, the number of genes in the top ranked list based on the systems biological analysis was not fixed by a preselected criterion.
The setup of the ORA is shown in Figure 4.1. The reference set consisted of all $m = 33239$ *Arabidopsis thaliana* genes listed in the TAIR database version 9. During the QTL study $n = 3133$ of these genes were identified as genes which are probably involved in biomass heterosis and we refer to these genes as the test set. A certain

number (x) of genes detected in our systems biological approach to heterosis with the highest Δs-values (Eq. 4.2 and 4.3) built the gene sets $C_{x,h,a}$ (separately for each hybrid $h \in \{\text{C24} \times \text{Col-0}, \text{Col-0} \times \text{C24}\}$ and heterosis measure $a \in \{\text{MPH, BPH}\}$). This led to four different gene sets for each x. The gene sets contained a maximum of $x = 8032$ genes in case all genes, even with negative Δs-values, were used. The number of genes in the reference set which belong to gene set $C_{x,h,a}$ is denoted by $l_{x,h,a}$. The number of genes of our test set which overlap with the current $C_{x,h,a}$ is given by $k_{x,h,a}$.

Given $l_{x,h,a}$, m and n, the number of genes ($k'_{x,h,a}$) that would be in the overlap between test set and gene set $C_{x,h,a}$ in the case that the test set is chosen randomly out of the reference set can be calculated:

$$k'_{x,h,a} = \frac{l_{x,h,a}}{m} * n \tag{4.4}$$

The genes of the gene set $C_{x,h,a}$ are called to be enriched in the test set if $k_{x,h,a}$ (genes observed in the overlap) is significantly larger than $k'_{x,h,a}$ (genes expected in the overlap just by chance) (Backes et al., 2007).

A hypergeometric distribution was used to estimate the probability of observing an overlap of $k_{x,h,a}$ genes between test set and gene set if these sets were independent (null hypothesis) (Drăghici et al., 2003; Fury et al., 2006). The probability (one-sided P-value) of having as many or more than $k_{x,h,a}$ genes in the overlap between test set and gene set $C_{x,h,a}$ can be calculated by summing up the probabilities of having $k_{x,h,a}$ or more genes belonging to the test set also in the gene set in the case that the genes of the test set are randomly chosen from the reference set (Drăghici et al., 2003):

$$P[X \geq k_{x,h,a} | m, n, l_{x,h,a}] = \sum_{i=k_{x,h,a}}^{n} \frac{\binom{l_{x,h,a}}{i}\binom{m-l_{x,h,a}}{n-i}}{\binom{m}{n}} \tag{4.5}$$

where X is a hypergeometric distributed random variable giving the size of the overlap. A P-value smaller than a given significance level corresponds to a significant over-representation of the gene set in the test set.

This whole analysis was performed in *R* (R Development Core Team, 2008). The function *phyper* was used to calculate the P-values corresponding to the hypergeometric distribution.

Fury et al. (2006) stated that the overlapping probability in an ORA changes with the number of genes in the test set and gene set. In our study the number of genes in the test set was fixed to 3133, but the number of genes in the gene set was not

fixed. Therefore, the ORA was run with different numbers of genes in the gene set to analyze the influence of the gene set size on the overlapping probability. For each run the genes with the x largest Δs-values were selected, separately for each hybrid as well as MPH and BPH, as the gene set. The number of genes in the gene set was ranged for the different ORA from $x = 0$ to $x = 8032$ (all genes in the systems biological network analysis) by steps of 100.

4.5.4 Resampling analysis of enrichment

In the ORA using a hypergeometric distribution, the probability to detect as many or more than the observed $k_{x,h,a}$ genes in the overlap between test set and gene set $C_{x,h,a}$ when a random test set is used is estimated. In a resampling analysis we calculated empirical resampling P-values on the basis of a resampling of the genes in the gene set. For each number x of genes in the gene sets we sampled new genes out of all genes in the reference set without replacement and assigned them to the original Δs-values. This is done 1000 times for any x. For each of these random gene sets the number of genes of the test set also present in the gene set were determined ($k^*_{x,h,a}$). The number of genes in the overlap between each original gene set and the test set is depicted by $k_{x,h,a}$. We used these values to calculate for each x the empirical resampling P-values:

$$P_{resampling,x,h,a} = \frac{\#(k^*_{x,h,a} \geq k_{x,h,a})}{\#resamplings} \tag{4.6}$$

Empirical resampling P-values ($P_{resampling,x,h,a}$) smaller than a given significance level are achieved in the case where the original gene set leads to a significantly larger overlap to the test set than expected by chance. This approach does not require distributional assumptions.

4.5.5 Chromosome-wise over-representation analysis

We analyzed if genes which are involved in biomass heterosis are functionally located at only some of the five *Arabidopsis thaliana* chromosomes.
In this study for any number x of genes in the "original" gene set, five ORA were performed. Each time the gene set $C_{x,h,a,chr}$ contained only the genes detected in our systems biological approach towards heterosis which belong to one of the five chromosomes ($chr \in \{1, ..., 5\}$). The assignment to the chromosomes was done based on the TAIR9 database. The reference set and test set of these ORA were the same as before.

Of all 8032 genes that were analyzed in the systems biological approach, 2105 belong to chromosome number 1, 1348 to chromosome 2, 1569 to chromosome 3, 1193 to chromosome 4 and 1807 to chromosome 5. 10 of the 8032 genes under study are not listed in the TAIR9 database.

4.5.6 Pathway analysis of candidate group of genes

In an attempt to get a little further insight into the molecular basis of biomass heterosis we determined the functional assignments of the genes in the overlap between the resulting genes from the systems biological analysis according to Andorf et al. (2010a) and the genes determined in the quantitative genetics approach by Meyer et al. (2010). This functional enrichment analysis was done by applying four further ORA. This time the reference set were all 8032 genes analyzed in our systems biological analysis. Four different test sets were used; one for each hybrid-heterosis measure combination. Each test set was built by first determining the 3000 genes with the highest Δs-values. Then the overlap of these 3000 genes to the 3133 genes detected in the quantitative genetics approach was identified and used as test set.
As gene sets we used 80 *Arabidopsis thaliana* pathways which contain between 10 and 4000 of the 8032 genes in the reference set. 28 of them were based on MapMan (Usadel et al., 2009), which in turn is based on the TAIR8 database. The remaining 52 pathways were built using Plant Ontology (PO) terms (The Plant Ontology Consortium, 2002).
The P-values achieved in this ORA were corrected for multiple testing using the FDR approach by Benjamini and Hochberg (1995).
In this setup we could determine if the overlapping genes between the two analyses are significantly enriched in one or more pathways (functional groups) of *Arabidopsis thaliana*.

Acknowledgments

This work was supported by the DFG (grants RE1654/2-1, SE611/3-1, AL387/6-1, AL387/6-2, AL387/6-3).

5 General discussion

Each chapter contains already a discussion about the particular results and the impact of the research described within it. Therefore, this general discussion will focus on the integration of the results of each separate chapter and will give an outlook of possible future research following the presented approach towards a better understanding of heterosis.

In this work, a systems biological approach towards biomass heterosis in *Arabidopsis thaliana* is presented to contribute to a better understanding of the heterosis phenomenon. The proposed network hypothesis for heterosis was tested on gene expression data as well as on metabolite profiles of two homozygous genotypes and their crosses. The estimated underlying regulatory networks of the homozygous parents and the hybrids were compared to reveal the differences in the network structure. Werhli et al. (2006) have shown that partial correlations of time series observational data can be used to estimate regulatory interactions. Hence, the determination of the differences in the regulatory networks was based on partial correlations according to Opgen-Rhein and Strimmer (2007b).

The network hypothesis for heterosis (section 1.7) predicted more regulatory interactions in the hybrids than in the homozygous genotypes. For the metabolite data, a high probability of more regulatory interactions in the hybrids compared to the parental lines was observed for MPH of both heterozygous genotypes (chapter 2). The hypothesis, tested on the metabolite profiles, was based on the partial correlation values. The difference of the mean partial correlation values of each metabolite between either hybrid and the mean of the parents was calculated.

A slightly different hypothesis was tested on the gene expression data (chapter 3). Instead of the partial correlation values itself, the significances (FDR corrected P-values) of the partial correlations built the basis for the calculation of the heterosis values. The hypothesis held true for either hybrid for MPH but only for C24 $\times$ Col-0 for the case of BPH.

To make a comparison between the results of these two omics datasets possible, the heterosis values of the metabolite data were, in an additional analysis, calculated on the basis of the significances of the partial correlations as in the analysis of the gene

expression profiles. Furthermore, different to the analysis in chapter 2, in which the hypothesis was only tested for MPH, the repeated analysis of the metabolite profile data was done for MPH as well as BPH. The results (appendix A on page 101) have shown that also the network hypothesis for heterosis based on the significance of partial correlations holds true for either hybrid and both heterosis measures for the experimental metabolite data in this work. A positive difference of the average significance of the partial correlations of the metabolites in the hybrids to the mean of the parents (for MPH) or to the better of the parents (for BPH) was observed for the majority of metabolites.

During the analysis based on the network hypothesis for heterosis, the cutoff, used in the filtering step after the linear model applied to the gene expression data (Eq. 3.3), to exclude genes from the further analysis that show nearly no time dependency and/or genotype-time point-interaction, proved to be a crucial factor. In a further analysis, the influence of this significance cutoff on the results of the analysis of the gene expression data regarding the network hypothesis was studied (appendix section B.1 on page 105). It was revealed that for each hybrid and heterosis measure, the use of small cutoffs for the FDR corrected P-values of the applied linear model (smaller than between 0.16 and 0.26, depending on the hybrid and heterosis measure) leave only genes for the heterosis analysis in this work that lead to rejection of the network hypothesis for heterosis. In case of "strict" filtering (small cutoffs) over all genotypes and time points, not only genes that show no time dependency and/or genotype-time point-interaction are excluded from the further analysis but also genes that led to weak time dependencies and/or genotype-time point-interactions. Conversely, for all cutoffs larger than these values, the hypothesis holds true for both hybrids and MPH as well as BPH.

Based on the analysis presented in appendix section B.1, the result from chapter 3, that the network hypothesis has to be rejected for Col-0 $\times$ C24 BPH, can be relativized. In chapter 3, a cutoff for the FDR corrected P-values was chosen that was in the range between 0.16 and 0.26, leading to the rejection of the hypothesis for Col-0 $\times$ C24 BPH. For every cutoff larger than 0.26, the analysis of the gene expression data presented in that chapter would have resulted in no rejection of the hypothesis for both hybrids and either heterosis measure.

The analysis of the influence of the significance cutoff on the rejection of the network hypothesis for heterosis was also tested for the metabolite data (appendix section B.2). The outcome was the same as for the gene expression profiles. For small significance cutoffs the hypothesis does not hold true for either hybrid regarding MPH as well as BPH. On the other side, for large cutoffs, the hypothesis is not

rejected for either hybrid and both heterosis measures, independent of the particular value of the cutoff.
Summarizing the results from chapters 2 and 3 and the in the appendix presented further analyses, for moderate large cutoffs applied to the FDR corrected P-values of the effects in the used linear models, the hypothesis of additional regulatory interactions in the heterozygous genotypes compared to the mid-parent and best-parent expectation is confirmed for metabolite as well as gene expression data for either hybrid regarding MPH and BPH.
In chapter 4, an over-representation analysis integrating the outcome from the application of the network hypothesis for heterosis to the gene expression profiles and the results from a quantitative genetics approach towards biomass heterosis in early *Arabidopsis thaliana* development (Meyer et al., 2010), was presented. A significant enrichment of the resulting genes of the systems biological analysis in the genes within the determined heterotic QTL regions was detected, leading to the suggestion that probably several genes in each region led to the detection of each heterotic QTL region.

In recent studies, the assumption was strengthened that the basic principle of heterosis is a complex interplay of several molecular mechanisms and based on the complex structure of biological networks, molecular regulation and inheritance processes (Birchler et al., 2003; Hochholdinger and Hoecker, 2007; Birchler et al., 2010). In heterosis analyses based on gene expression data of different inbred lines and crosses, no consensus set of genes could be identified that was differentially expressed between all inbred-hybrid combinations. Hochholdinger and Hoecker (2007) concluded that, since no key genes for heterosis could be found in these studies, rather *global* trends of gene expression are correlated with heterosis.
Therefore, we are convinced that systems biological approaches towards the molecular basis of heterosis, such as the one presented in this work, are promising to contribute to a better understanding of the heterosis phenomenon. These approaches should not be seen as an alternative to quantitative genetics analyses but as complementary.
Up to now, heterosis analyses based on high-throughput data almost always followed the aim to identify pathways containing genes that influence the heterosis phenotype or, in the molecular biological view, to determine and characterize single genes that have an impact on the phenotype under study. The here presented analysis, based on the network hypothesis for heterosis, is to our knowledge one of the first approaches that studies all active parts of the regulatory networks of different genotypes with

respect to heterosis. Our approach is independent of any known biological function of the genes and metabolites. Only the changes in the regulatory network structure are analyzed. This is in contrast to other high-throughput molecular biological approaches which are focusing on finding responsible biological pathways. This basic research approach may deliver a valuable basis for a further understanding of the basis of heterosis.

The network hypothesis for heterosis was tested and confirmed on data of two omics levels of the same *Arabidopsis thaliana* genotypes. The time points during the development at which the samples were taken are not exactly the same for the metabolomic and transcriptomic data. However, in both experiments, time points during the early development of *Arabidopsis thaliana* were analyzed, making a comparison of the results permissible. Hence, this study of the same biological question on different levels of biological organization can bee seen as an integrative approach. The conclusion that the hybrids contain denser regulatory networks than the homozygous parental lines is strengthened, in comparison to the analysis of only one dataset, by the fact that experimental data of two different omics levels have shown the same outcome.

Other approaches to integrate gene expression and metabolite data are presented in the literature. For example, the integrative systems biological analysis of metabolite and gene expression data by Urbanczyk-Wochniak et al. (2003) is based on a pairwise correlation analysis between metabolite and gene expression data. For each transcript it is determined whether it is correlated with any of the metabolites under study. They conclude that this integrative analysis can be used for a rapid identification of candidate genes. In the analysis by Urbanczyk-Wochniak et al. (2003), the metabolite and gene expression data were measured from the same samples at the same time points. However, as explained above, for the heterosis analysis presented here, the metabolite and gene expression data were measured at different time points during the early development of *Arabidopsis thaliana*. Therefore, as promising as the approach by Urbanczyk-Wochniak et al. (2003) is, it can not be applied straight forward to the experimental datasets in this work.

As yet, reverse engineering approaches are widely used for transcriptomics data, but the application on metabolomics is rather limited (Çakır et al., 2009). In chapter 2, a reverse engineering analysis applied to metabolite data was presented. Since both reverse engineering approaches, applied to metabolomic and transcriptomic data, came to the same conclusion regarding the network hypothesis for heterosis, we could hypothize that, even though it was not often used so far, reverse engineering approaches based on partial correlations can be used to infer information about

regulatory structures of metabolite networks. When we performed this analysis on the metabolite data, no comparative study has been done to study the applicability of this approach to reconstruct metabolic networks. By now, Çakır et al. (2009) performed such a study and could show that reverse engineering approaches based on partial correlations can indeed infer, up to a certain accuracy, metabolomic networks from observational steady state metabolite data.

Reverse engineering approaches are rather applicable if the measured samples contain only one type of cells than a heterogeneous mixture of different cell types (Venet et al., 2001; Lu et al., 2003). The basis for the metabolite as well as the gene expression data in this work were the whole *Arabidopsis thaliana* seedlings. Due to the fact that the heterozygous genotypes show biomass heterosis, the mixture of cell types most probably differs at any given time point between the genetically distinct plants. This different composition of cell types might have an interfering effect on the outcome of the reverse engineering analyses in this work. However, we assume that this effect is smaller than the actual effect caused by the different regulatory networks that we want to study and, hence, it does not influence the reliability of the global conclusion about more regulatory interactions.

The integrative analysis for the results of the systems biological approach and the heterotic QTL regions identified by Meyer et al. (2010) revealed a significantly larger overlap between the two approaches than expected by chance. Instead of an ORA, this analysis could have been performed based on a GSEA. As described in the introduction (section 1.6) GSEA use lists which are sorted by some criterion as test set. The GSEA is, therefore, not applicable straight forward in the setup presented in chapter 4 because the genes within the heterotic QTL regions, that built the test set, were not quantitative data. The quantitative character of these genes, however, could be achieved by using the distance of each gene to the closest peak of the LOD score profile (personal communication, Prof. T. Altmann, IPK Gatersleben).

In addition, in chapter 4, an over-representation of genes from biomass related *Arabidopsis thaliana* pathways in the overlapping genes between the both integrated approaches was observed. Since both approaches analyzed biomass heterosis in the early development of *Arabidopsis thaliana*, the enrichment in these pathways confirms that both approaches pointed to similar sets of genes influencing biomass heterosis in *Arabidopsis thaliana*. This conclusion goes along with the other result of this integrative approach that not only a few but several genes influencing biomass heterosis are located within each heterotic QTL region.

A first extension of the study of the network hypothesis for heterosis in *Arabidopsis thaliana* could be based on the work by Lisec et al. (2009). They analyzed two mapping populations of a cross between the same two *Arabidopsis thaliana* accessions C24 and Col-0, as used in this work, regarding heterosis at the metabolic level. Including their identified heterotic metabolic QTL, would complete the "set" of analyses towards biomass heterosis in *Arabidopsis thaliana* in this work a little further. This "set" would then contain the results of the systems biological approach based on the network hypothesis for heterosis for the metabolite as well as gene expression data and the outcome of QTL mapping approaches towards biomass heterosis for metabolomic (Lisec et al., 2009) as well as for transcriptomic data (Meyer et al., 2010). Based on this set, comparisons and integrations of the results of the same approach applied to different omics levels, results of different approaches but used for the same omics level or any other combination of these approaches and omics levels would be possible. Integrative analyses of these four analyses (two approaches $\times$ two omics levels) towards biomass heterosis in *Arabidopsis thaliana* would account for the limitations, restrictions and characteristics of either analysis method and either omics level. This, in turn, would be a further step towards a better understanding of the molecular basis of the heterosis phenomenon.

One important point in the evaluation of this work is that all the results can not be used to draw general conclusions. The reason for this is the small experimental data basis. Only two homozygous lines and the reciprocal crosses of one species (*Arabidopsis thaliana*) were analyzed for one heterosis trait. Furthermore, even though it was shown in chapter 3 that it is possible to infer biological networks from the experimental data of seven time points, this data basis is still too weak to draw general conclusions. As pointed out by Hochholdinger and Hoecker (2007), the outcome of heterosis analyses, such as which model (dominance, overdominance or epistasis) is the most favorable, are controversy and depend up to some extend e.g. on the analyzed organism, type of tissue, developmental stage or the experimental technique that is used for the analysis. Hence, to draw more general conclusions about the molecular processes underlying heterosis, based on the presented network hypothesis for heterosis, this hypothesis should be tested on experimental data from more species, different developmental stages and diverse heterosis traits.

Summarizing, the systems biological approach towards the better understanding of heterosis presented in this work, is one of the first analyses based on changes in all active parts of the regulatory networks of homozygous to heterozygous genotypes (chapters 2 and 3). On two different omics levels it was estimated that the heterozy-

gous *Arabidopsis thaliana* genotypes contain denser regulatory networks than the homozygous parents. The results of this systems biological approach were integrated with QTL mapping experiments towards biomass heterosis, resulting in a significant overlap between the results of the two different analyses (chapter 4). This led to the suggestion that each heterotic QTL region contains many genes probably influencing biomass heterosis in early *Arabidopsis thaliana* development.

References

Ackermann, M. and Strimmer, K. (2009). A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, 10:47.

Agilent Technologies Inc. (2008). *Agilent Feature Extraction Software - Reference Guide*. USA, G4460-90020, sixth edition.

Andorf, S., Gärtner, T., Steinfath, M., Witucka-Wall, H., Altmann, T., and Repsilber, D. (2009). Towards systems biology of heterosis: A hypothesis about molecular network structure applied for the arabidopsis metabolome. *EURASIP J Bioinform Syst Biol*, 2009:articleID: 147157. Special Issue: Network structure and biological function: reconstruction, modelling, and statistical approaches.

Andorf, S., Selbig, J., Altmann, T., Poos, K., Witucka-Wall, H., and Repsilber, D. (2010a). Enriched partial correlations in genome-wide gene expression profiles of hybrids (A. thaliana): a systems biological approach towards the molecular basis of heterosis. *Theor Appl Genet*, 120(2):249–259.

Andorf, S., Selbig, J., Altmann, T., Witucka-Wall, H., and Repsilber, D. (2010b). Heterosis in arabidopsis thaliana: A metabolite network structure approach. *Schriftenreihe des Leibniz-Instituts für Nutztierbiologie*, 16:7–10.

Backes, C., Keller, A., Kuentzer, J., Kneissl, B., Comtesse, N., Elnakady, Y. A., Müller, R., Meese, E., and Lenhof, H.-P. (2007). GeneTrail–advanced gene set enrichment analysis. *Nucleic Acids Res*, 35(Web Server issue):W186–W192.

Barabási, A. L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.

Barabási, A.-L. and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat Rev Genet*, 5(2):101–113.

Basso, K., Margolin, A. A., Stolovitzky, G., Klein, U., Dalla-Favera, R., and Califano, A. (2005). Reverse engineering of regulatory networks in human b cells. *Nat Genet*, 37(4):382–390.

Beal, M. J., Falciani, F., Ghahramani, Z., Rangel, C., and Wild, D. L. (2005). A bayesian approach to reconstructing genetic regulatory networks with hidden factors. *Bioinformatics*, 21(3):349–356.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Statist Soc B*, 57(1):289–300.

Birchler, J. A., Auger, D. L., and Riddle, N. C. (2003). In search of the molecular basis of heterosis. *Plant Cell*, 15(10):2236–2239.

Birchler, J. A., Yao, H., Chudalayandi, S., Vaiman, D., and Veitia, R. A. (2010). Heterosis. *Plant Cell*, 22(7):2105–2112.

Bohm, D. (1980). *Wholeness And The Implicate Order*. Routledge, London.

Brazhnik, P., de la Fuente, A., and Mendes, P. (2002). Gene networks: how to put the function in genomics. *Trends Biotechnol*, 20(11):467–472.

Bruce, A. B. (1910). The mendelian theory of heredity and the augmentation of vigor. *Science*, 32(827):627–628.

Butte, A. J., Tamayo, P., Slonim, D., Golub, T. R., and Kohane, I. S. (2000). Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc Natl Acad Sci U S A*, 97(22):12182–12186.

Campbell, A. M. and Heyer, L. J. (2002). *Discovering Genomics, Proteomics and Bioinformatics*. Benjamin Cummings.

Çakır, T., Hendriks, M. M. W. B., Westerhuis, J. A., and Smilde, A. K. (2009). Metabolic network discovery through reverse engineering of metabolome data. *Metabolomics*, 5(3):318–329.

Churchill, G. A. and Doerge, R. W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics*, 138(3):963–971.

Crow, J. F. (1948). Alternative hypotheses of hybrid vigor. *Genetics*, 33(5):477–487.

Crow, J. F. (1952). *Heterosis*, chapter Dominance and Overdominance, pages 282–297. Iowa State College Press, Ames, IA.

Cui, Q., Lewis, I. A., Hegeman, A. D., Anderson, M. E., Li, J., Schulte, C. F., Westler, W. M., Eghbalnia, H. R., Sussman, M. R., and Markley, J. L. (2008). Metabolite identification via the madison metabolomics consortium database. *Nat Biotechnol*, 26(2):162–164.

Davenport, C. B. (1908). Degeneration, albinism and inbreeding. *Science*, 28(718):454–455.

de Jong, H. (2002). Modeling and simulation of genetic regulatory systems: a literature review. *J Comput Biol*, 9(1):67–103.

de la Fuente, A., Bing, N., Hoeschele, I., and Mendes, P. (2004). Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics*, 20(18):3565–3574.

de Silva, E., Thorne, T., Ingram, P., Agrafioti, I., Swire, J., Wiuf, C., and Stumpf, M. P. H. (2006). The effects of incomplete protein interaction data on structural and evolutionary inferences. *BMC Biol*, 4:39.

D'haeseleer, P., Wen, X., Fuhrman, S., and Somogyi, R. (1999). Linear modeling of mrna expression levels during cns development and injury. *Pac Symp Biocomput*, pages 41–52.

Doniger, S. W., Salomonis, N., Dahlquist, K. D., Vranizan, K., Lawlor, S. C., and Conklin, B. R. (2003). MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol*, 4(1):R7.

Drăghici, S., Khatri, P., Martins, R. P., Ostermeier, G. C., and Krawetz, S. A. (2003). Global functional profiling of gene expression. *Genomics*, 81(2):98–104.

East, E. M. (1936). Heterosis. *Genetics*, 21(4):375–397.

Falconer, D. S. and Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics*. Longman, Essex, England.

Fiehn, O., Kopka, J., Drmann, P., Altmann, T., Trethewey, R. N., and Willmitzer, L. (2000). Metabolite profiling for plant functional genomics. *Nat Biotechnol*, 18(11):1157–1161.

Frascaroli, E., Can, M. A., Landi, P., Pea, G., Gianfranceschi, L., Villa, M., Morgante, M., and P, M. E. (2007). Classical genetic and quantitative trait loci analyses of heterosis in a maize hybrid between two elite inbred lines. *Genetics*, 176(1):625–644.

Friedman, J. H. (1989). Regularized discriminant analysis. *J Am Stat Assoc*, 84:165–175.

Friedman, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science*, 303(5659):799–805.

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using bayesian networks to analyze expression data. *J Comput Biol*, 7(3-4):601–620.

Frisch, M., Thiemann, A., Fu, J., Schrag, T. A., Scholten, S., and Melchinger, A. E. (2010). Transcriptome-based distance measures for grouping of germplasm and prediction of hybrid performance in maize. *Theor Appl Genet*, 120(2):441–450.

Fury, W., Batliwalla, F., Gregersen, P. K., and Li, W. (2006). Overlapping probabilities of top ranking gene lists, hypergeometric distribution, and stringency of gene selection criterion. *Conf Proc IEEE Eng Med Biol Soc*, 1:5531–5534.

Gardner, T. S. and Faith, J. J. (2005). Reverse-engineering transcription control networks. *Phys Life Rev*, 2(1):65–88.

Gärtner, T., Steinfath, M., Andorf, S., Lisec, J., Meyer, R. C., Altmann, T., Willmitzer, L., and Selbig, J. (2009). Improved heterosis prediction by combining information on dna- and metabolic markers. *PLoS One*, 4(4):e5220.

Ge, H., Walhout, A. J. M., and Vidal, M. (2003). Integrating 'omic' information: a bridge between genomics and systems biology. *Trends Genet*, 19(10):551–560.

Genoud, T. and Métraux, J.-P. (1999). Crosstalk in plant cell signaling: structure and function of the genetic network. *Trends Plant Sci*, 4(12):503–507.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y., and Zhang, J. (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol*, 5:R80.

Gibson, G. (1996). Epistasis and pleiotropy as natural properties of transcriptional regulation. *Theor Popul Biol*, 49(1):58–89.

Gjuvsland, A. B., Hayes, B. J., Omholt, S. W., and Carlborg, O. (2007). Statistical epistasis is a generic feature of gene regulatory networks. *Genetics*, 175(1):411–420.

Goeman, J. J. and Bühlmann, P. (2007). Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980–987.

Greenland, S. (2000). Principles of multilevel modelling. *Int J Epidemiol*, 29(1):158–167.

Guimerà, R. and Amaral, L. A. N. (2005). Functional cartography of complex metabolic networks. *Nature*, 433(7028):895–900.

Guo, M., Rupe, M. A., Yang, X., Crasta, O., Zinselmeier, C., Smith, O. S., and Bowen, B. (2006). Genome-wide transcript analysis of maize hybrids: allelic additive gene expression and yield heterosis. *Theor Appl Genet*, 113(5):831–845.

Hache, H., Lehrach, H., and Herwig, R. (2009). Reverse engineering of gene regulatory networks: a comparative study. *EURASIP J Bioinform Syst Biol*, 2009:articleID: 617281.

Hackett, C. A. (2002). Statistical methods for QTL mapping in cereals. *Plant Mol Biol*, 48(5-6):585–599.

Harrison, G. A. (1962). Heterosis and adaptability in the heat tolerance of mice. *Genetics*, 47(4):427–434.

Hartemink, A. J. (2005). Reverse engineering gene regulatory networks. *Nat Biotechnol*, 23(5):554–555.

Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. (1999). From molecular to modular cell biology. *Nature*, 402(SUPP):C47–C52.

Hochholdinger, F. and Hoecker, N. (2007). Towards the molecular basis of heterosis. *Trends Plant Sci*, 12(9):427–432.

Hood, L. and Galas, D. (2003). The digital code of dna. *Nature*, 421(6921):444–448.

Huala, E., Dickerman, A. W., Garcia-Hernandez, M., Weems, D., Reiser, L., LaFond, F., Hanley, D., Kiphart, D., Zhuang, M., Huang, W., Mueller, L. A., Bhattacharyya, D., Bhaya, D., Sobral, B. W., Beavis, W., Meinke, D. W., Town, C. D., Somerville, C., and Rhee, S. Y. (2001). The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Res*, 29(1):102–105.

Hull, F. H. (1945). Recurrent selection for specific combining ability in corn. *Agron J*, 37:134–145.

Husmeier, D. (2003). Reverse engineering of genetic networks with bayesian networks. *Biochem Soc Trans*, 31(Pt 6):1516–1518.

Ideker, T., Thorsson, V., Ranish, J. A., Christmas, R., Buhler, J., Eng, J. K., Bumgarner, R., Goodlett, D. R., Aebersold, R., and Hood, L. (2001). Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, 292(5518):929–934.

Imoto, S., Higuchi, T., Goto, T., Tashiro, K., Kuhara, S., and Miyano, S. (2003). Combining microarrays and biological knowledge for estimating gene networks via bayesian networks. *Proc IEEE Comput Soc Bioinform Conf*, 2:104–113.

Ingolia, N. T. and Weissman, J. S. (2008). Systems biology: Reverse engineering the cell. *Nature*, 454(7208):1059–1062.

Kanehisa, M. and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28(1):27–30.

Kauffman, S. A. (1993). *The Origins of Order*. University Press, Oxford.

Kell, D. B. (2004). Metabolomics and systems biology: making sense of the soup. *Curr Opin Microbiol*, 7(3):296–307.

Kell, D. B., Brown, M., Davey, H. M., Dunn, W. B., Spasic, I., and Oliver, S. G. (2005). Metabolic footprinting and systems biology: the medium is the message. *Nat Rev Microbiol*, 3(7):557–565.

Kerr, M. K. and Churchill, G. A. (2001). Experimental design for gene expression microarrays. *Biostatistics*, 2(2):183–201.

Kerr, M. K., Martin, M., and Churchill, G. A. (2000). Analysis of variance for gene expression microarray data. *J Comput Biol*, 7(6):819–837.

Khatri, P. and Draghici, S. (2005). Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, 21(18):3587–3595.

Kishino, H. and Waddell, P. J. (2000). Correspondence analysis of genes and tissue types and finding genetic links from microarray data. *Genome Informatics*, 11:83–95.

Kitano, H. (2002). Systems biology: a brief overview. *Science*, 295(5560):1662–1664.

Lachmann, A. and Ma'ayan, A. (2010). Lists2networks: integrated analysis of gene/protein lists. *BMC Bioinformatics*, 11:87.

Lambert, D. and Hughes, T. (1984). Misery of functionalism. Biological function: a misleading concept. *Riv Biol*, 77(4):477–502.

Lamkey, K. R. and Edwards, J. W. (1999). The quantitative genetics of heterosis. In Coors, J. and Pandey, S., editors, *The Genetics and Exploitation of Heterosis in Crops*, pages 31–48. ASA, CSSA, and SSSA, Madison, WI.

Ledoit, O. and Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *J Empir Finance*, 10:603–621.

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., Zeitlinger, J., Jennings, E. G., Murray, H. L., Gordon, D. B., Ren, B., Wyrick, J. J., Tagne, J.-B., Volkert, T. L., Fraenkel, E., Gifford, D. K., and Young, R. A. (2002). Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, 298(5594):799–804.

Li, Z. K., Luo, L. J., Mei, H. W., Wang, D. L., Shu, Q. Y., Tabien, R., Zhong, D. B., Ying, C. S., Stansel, J. W., Khush, G. S., and Paterson, A. H. (2001). Overdominant epistatic loci are the primary genetic basis of inbreeding depression and heterosis in rice. I. biomass and grain yield. *Genetics*, 158(4):1737–1753.

Link, W., Schill, B., Barbera, A. C., Cubero, J. I., Filippetti, A., Stringi, L., von E. Kittlitz, and Melchinger, A. E. (1996). Comparison of intra- and inter-pool crosses in faba beans (*vicia faba* L.). I. Hybrid performance and heterosis in Mediterranean and German environments. *Plant Breeding*, 115(5):352–360.

Lisec, J., Schauer, N., Kopka, J., Willmitzer, L., and Fernie, A. R. (2006). Gas chromatography mass spectrometry-based metabolite profiling in plants. *Nat Protoc*, 1(1):387–396.

Lisec, J., Steinfath, M., Meyer, R. C., Selbig, J., Melchinger, A. E., Willmitzer, L., and Altmann, T. (2009). Identification of heterotic metabolite qtl in arabidopsis thaliana ril and il populations. *Plant J*, 59(5):777–788.

Lu, P., Nakorchevskiy, A., and Marcotte, E. M. (2003). Expression deconvolution: a reinterpretation of dna microarray data reveals dynamic changes in cell populations. *Proc Natl Acad Sci U S A*, 100:10370–10375.

Luo, L. J., Li, Z. K., Mei, H. W., Shu, Q. Y., Tabien, R., Zhong, D. B., Ying, C. S., Stansel, J. W., Khush, G. S., and Paterson, A. H. (2001). Overdominant epistatic loci are the primary genetic basis of inbreeding depression and heterosis in rice. II. grain yield components. *Genetics*, 158(4):1755–1771.

Ma, L., Sun, N., Liu, X., Jiao, Y., Zhao, H., and Deng, X. W. (2005). Organ-specific expression of Arabidopsis genome during development. *Plant Physiol*, 138(1):80–91.

Ma, S., Gong, Q., and Bohnert, H. J. (2007). An arabidopsis gene network based on the graphical gaussian model. *Genome Res*, 17(11):1614–1625.

Magwene, P. M. and Kim, J. (2004). Estimating genomic coexpression networks using first-order conditional independence. *Genome Biol*, 5(12):R100.

Martínez, P., López, C., Roldán, M., Sabater, B., and Martín, M. (1997). Plastid DNA of five ecotypes of Arabidopsis thaliana: sequence of ndhG gene and maternal inheritance. *Plant Sci*, 123(1-2):113–122.

Matthäus, F., Salazar, C., and Ebenhöh, O. (2008). Biosynthetic potentials of metabolites and their hierarchical organization. *PLoS Comput Biol*, 4(4):e1000049.

Maynard Smith, J. (1956). Acclimatization to high temperatures in inbred and outbred Drosophila subobscura. *J. Genet.*, 54(1):497–505.

Melchinger, A. E. (1999). *The Genetics and Exploitation of Heterosis in Crops*, chapter Genetic diversity and heterosis, pages 99–118. ASA-CSSA, Madison WI, USA.

Melchinger, A. E., Piepho, H.-P., Utz, H. F., Muminovic, J., Wegenast, T., Trjk, O., Altmann, T., and Kusterer, B. (2007a). Genetic basis of heterosis for growth-related traits in arabidopsis investigated by testcross progenies of near-isogenic lines reveals a significant role of epistasis. *Genetics*, 177(3):1827–1837.

Melchinger, A. E., Utz, H. F., Piepho, H.-P., Zeng, Z.-B., and Schön, C. C. (2007b). The role of epistasis in the manifestation of heterosis: a systems-oriented approach. *Genetics*, 177(3):1815–1825.

Meyer, R. C., Kusterer, B., Lisec, J., Steinfath, M., Becher, M., Scharr, H., Melchinger, A. E., Selbig, J., Schurr, U., Willmitzer, L., and Altmann, T. (2010). QTL analysis of early stage heterosis for biomass in Arabidopsis. *Theor Appl Genet*, 120(2):227–237.

Meyer, R. C., Törjék, O., Becher, M., and Altmann, T. (2004). Heterosis of biomass production in Arabidopsis. establishment during early development. *Plant Physiol*, 134(4):1813–1823.

Meyer, S., Pospisil, H., and Scholten, S. (2007). Heterosis associated gene expression in maize embryos 6 days after fertilization exhibits additive, dominant and overdominant pattern. *Plant Mol Biol*, 63(3):381–391.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827.

Moll, R. H., Lonnquist, J. H., Fortuno, J. V., and Johnson, E. C. (1965). The relationship of heterosis and genetic divergence in maize. *Genetics*, 52(1):139–144.

Mootha, V. K., Lindgren, C. M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., Houstis, N., Daly, M. J., Patterson, N., Mesirov, J. P., Golub, T. R., Tamayo, P., Spiegelman, B., Lander, E. S., Hirschhorn, J. N., Altshuler, D., and Groop, L. C. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet*, 34(3):267–273.

Mount, D. W. (2004). *Bioinformatics: Sequence and Genome Analysis*. Cold Spring Harbor Laboratory Press, 2nd edition.

Nadal, J.-P. (1991). Associative memory: on the (puzzling) sparse coding limits. *Journal of Physics A: Mathematical and General*, 24(5):1093–1101.

Noble, D. (2002). Modeling the heart–from genes to cells to the whole organ. *Science*, 295(5560):1678–1682.

Oliver, S. G., Winson, M. K., Kell, D. B., and Baganz, F. (1998). Systematic functional analysis of the yeast genome. *Trends Biotechnol*, 16(9):373–378.

Opgen-Rhein, R., Schäfer, J., and Strimmer, K. (2007). *GeneNet: Modeling and Inferring Gene Networks*. R package version 1.2.0.

Opgen-Rhein, R. and Strimmer, K. (2007a). Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach. *Stat Appl Genet Mol Biol*, 6:Article9.

Opgen-Rhein, R. and Strimmer, K. (2007b). From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC Syst Biol*, 1:37.

Parrish, R. S., Spencer III, H. J., and Xu, P. (2009). Distribution modeling and simulation of gene expression data. *Comput Stat Data Anal*, 53(5):1650–1660.

Perrot, M., Guieysse-Peugeot, A.-L., Massoni, A., Espagne, C., Claverol, S., Silva, R. M., Jenö, P., Santos, M., Bonneu, M., and Boucherie, H. (2007). Yeast proteome map (update 2006). *Proteomics*, 7(7):1117–1120.

Powers, L. (1944). An expansion of jone's theory for the explanation of heterosis. *Am Nat*, 78:275–280.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Riley, M. C., Clare, A., and King, R. D. (2007). Locational distribution of gene functional classes in Arabidopsis thaliana. *BMC Bioinformatics*, 8:112.

Ritchie, M. E., Silver, J., Oshlack, A., Holmes, M., Diyagama, D., Holloway, A., and Smyth, G. K. (2007). A comparison of background correction methods for two-colour microarrays. *Bioinformatics*, 23:2700–2707.

Rivals, I., Personnaz, L., Taing, L., and Potier, M.-C. (2007). Enrichment or depletion of a go category within a class of genes: which test? *Bioinformatics*, 23(4):401–407.

Robertson, F. W. and Reeve, E. C. (1952). Heterozygosity, environmental variation and heterosis. *Nature*, 170(4320):286.

Ruf, S., Karcher, D., and Bock, R. (2007). Determining the transgene containment level provided by chloroplast transformation. *Proc Natl Acad Sci U S A*, 104(17):6998–7002.

Saul, Z. M. and Filkov, V. (2007). Exploring biological network structure using exponential random graph models. *Bioinformatics*, 23(19):2604–2611.

Schäfer, J., Opgen-Rhein, R., , and Strimmer, K. (2009). *GeneNet: Modeling and Inferring Gene Networks*. R package version 1.2.4 http://CRAN.R-project.org/package=GeneNet.

Schäfer, J., Opgen-Rhein, R., and Strimmer, K. (2006). Reverse engineering genetic networks using the "genenet" package. *R News*, 6/5:50–53.

Schäfer, J. and Strimmer, K. (2005a). An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21(6):754–764.

Schäfer, J. and Strimmer, K. (2005b). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat Appl Genet Mol Biol*, 4:Article32.

Schnell, F. W. and Cockerham, C. C. (1992). Multiplicative vs. arbitrary gene action in heterosis. *Genetics*, 131(2):461–469.

Schrag, T. A., Mhring, J., Melchinger, A. E., Kusterer, B., Dhillon, B. S., Piepho, H.-P., and Frisch, M. (2010). Prediction of hybrid performance in maize using molecular markers and joint analyses of hybrids and parental inbreds. *Theor Appl Genet*, 120(2):451–461.

Shen-Orr, S. S., Milo, R., Mangan, S., and Alon, U. (2002). Network motifs in the transcriptional regulation network of Escherichia coli. *Nat Genet*, 31(1):64–68.

Shubik, M. (1996). Simulations, models and simplicity. *Complexity*, 2(1):60.

Shull, G. H. (1908). The composition of a field of maize. *Am Breeders Assoc Rep*, 4:296–301.

Shull, G. H. (1948). What is "heterosis"? *Genetics*, 33(5):439–446.

Shull, G. H. (1952). Beginnings of the heterosis concept. In Gowen, J. W., editor, *Heterosis: a record of researches directed toward explaining and utilizing the vigor of hybrids*, pages 14–48. Iowa State College Press, Ames.

Smyth, G. K. (2005). Limma: linear models for microarray data. In Gentleman, R., Carey, V., Dudoit, S., Irizarry, R., and Huber, W., editors, *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, pages 397–420. Springer, New York.

Smyth, G. K. and Speed, T. (2003). Normalization of cDNA microarray data. *Methods*, 31:265–273.

Solomon, K. F., Labuschagne, M. T., and Viljoen, C. D. (2007). Estimates of heterosis and association of genetic distance with heterosis in durum wheat under different moisture conditions. *Journal of Agricultural science*, 145:239–248.

Somogyi, R. and Sniegoski, C. A. (1996). Modeling the complexity of genetic networks: understanding multigenic and pleiotropic regulation. *Complexity*, 1:45–63.

Song, R. and Messing, J. (2003). Gene expression of a gene family in maize based on noncollinear haplotypes. *Proc Natl Acad Sci U S A*, 100(15):9055–9060.

Steinbuch, K. (1961). Die Lernmatrix. *Kybernetik*, 1:36–45.

Steinfath, M., Gärtner, T., Lisec, J., Meyer, R. C., Altmann, T., Willmitzer, L., and Selbig, J. (2010). Prediction of hybrid biomass in Arabidopsis thaliana by selected parental SNP and metabolic markers. *Theor Appl Genet*, 120(2):239–247.

Steinfath, M., Repsilber, D., Scholz, M., Walther, D., and Selbig, J. (2007). Integrated data analysis for genome-wide research. In Baginsky, S. and Fernie, A. R., editors, *Plant Systems Biology*, volume 97 of *Experientia Supplementum*, pages 309–329. Birkhäuser Basel.

Strimmer, K. (2008). A unified approach to false discovery rate estimation. *BMC Bioinformatics*, 9:303.

Strogatz, S. H. (2001). Exploring complex networks. *Nature*, 410(6825):268–276.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102(43):15545–15550.

Swanson-Wagner, R. A., Jia, Y., DeCook, R., Borsuk, L. A., Nettleton, D., and Schnable, P. S. (2006). All possible modes of gene action are observed in a global comparison of gene expression in a maize F1 hybrid and its inbred parents. *Proc Natl Acad Sci U S A*, 103(18):6805–6810.

Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T. Z., Garcia-Hernandez, M., Foerster, H., Li, D., Meyer, T., Muller, R., Ploetz, L., Radenbaugh, A., Singh, S., Swing, V., Tissier, C., Zhang, P., and Huala, E. (2008). The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.*, 36(Database issue):D1009–D1014.

The Plant Ontology Consortium (2002). The plant ontology consortium and plant ontologies. *Comp Funct Genomics*, 3:137–142.

Thiemann, A., Fu, J., Schrag, T. A., Melchinger, A. E., Frisch, M., and Scholten, S. (2010). Correlation between parental transcriptome and field data for the characterization of heterosis in Zea mays L. *Theor Appl Genet*, 120(2):401–413.

Thimm, O., Bläsing, O., Gibon, Y., Nagel, A., Meyer, S., Krüger, P., Selbig, J., Müller, L. A., Rhee, S. Y., and Stitt, M. (2004). MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J*, 37(6):914–939.

Toh, H. and Horimoto, K. (2002). Inference of a genetic network by a combined approach of cluster analysis and graphical gaussian modeling. *Bioinformatics*, 18(2):287–297.

Törjék, O., Witucka-Wall, H., Meyer, R. C., von Korff, M., Kusterer, B., Rautengarten, C., and Altmann, T. (2006). Segregation distortion in Arabidopsis C24/Col-0 and Col-0/C24 recombinant inbred line populations is due to reduced fertility caused by epistatic interaction of two loci. *Theor Appl Genet*, 113(8):1551–1561.

Tresch, A. and Markowetz, F. (2008). Structure learning in nested effects models. *Stat Appl Genet Mol Biol*, 7(1):Article9.

Tsaftaris, S. A. (1995). Molecular aspects of heterosis in plants. *Physiologia Plantarum*, 94(2):362–370.

Tsamardinos, I., Brown, L. E., and Aliferis, C. F. (2006). The max-min hill-climbing Bayesian network structure learning algorithm. *Mach Learn*, 65:31–78.

Urbanczyk-Wochniak, E., Luedemann, A., Kopka, J., Selbig, J., Roessner-Tunali, U., Willmitzer, L., and Fernie, A. R. (2003). Parallel analysis of transcript and metabolic profiles: a new approach in systems biology. *EMBO Rep*, 4(10):989–993.

Usadel, B., Poree, F., Nagel, A., Lohse, M., Czedik-Eysenberg, A., and Stitt, M. (2009). A guide to using MapMan to visualize and compare Omics data in plants: a case study in the crop species, Maize. *Plant Cell Environ*, 32(9):1211–1229.

Utz, H. F. and Melchinger, A. E. (1996). PLABQTL: A program for composite interval mapping of QTL. *J QTL*, 2.

Velculescu, V. E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M. A., Bassett, D. E., Hieter, P., Vogelstein, B., and Kinzler, K. W. (1997). Characterization of the yeast transcriptome. *Cell*, 88(2):243–251.

Venet, D., Pecasse, F., Maenhaut, C., and Bersini, H. (2001). Separation of samples into their constituents using gene expression data. *Bioinformatics*, 17 Suppl 1:S279–S287.

Vuylsteke, M., van Eeuwijk, F., Hummelen, P. V., Kuiper, M., and Zabeau, M. (2005). Genetic analysis of variation in gene expression in Arabidopsis thaliana. *Genetics*, 171(3):1267–1275.

Wang, Z., Gerstein, M., and Snyder, M. (2009). Rna-seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1):57–63.

Weckwerth, W. (2003). Metabolomics in systems biology. *Annu Rev Plant Biol*, 54:669–689.

Wei, G., Tao, Y., Liu, G., Chen, C., Luo, R., Xia, H., Gan, Q., Zeng, H., Lu, Z., Han, Y., Li, X., Song, G., Zhai, H., Peng, Y., Li, D., Xu, H., Wei, X., Cao, M., Deng, H., Xin, Y., Fu, X., Yuan, L., Yu, J., Zhu, Z., and Zhu, L. (2009). A transcriptomic analysis of superhybrid rice lyp9 and its parents. *Proc Natl Acad Sci U S A*, 106(19):7695–7701.

Wei, H., Persson, S., Mehta, T., Srinivasasainagendra, V., Chen, L., Page, G. P., Somerville, C., and Loraine, A. (2006). Transcriptional coordination of the metabolic network in arabidopsis. *Plant Physiol*, 142(2):762–774.

Werhli, A. V., Grzegorczyk, M., and Husmeier, D. (2006). Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. *Bioinformatics*, 22(20):2523–2531.

Westerhoff, H. V. and Palsson, B. O. (2004). The evolution of molecular biology into systems biology. *Nat Biotechnol*, 22(10):1249–1252.

Wille, A., Zimmermann, P., Vranová, E., Frholz, A., Laule, O., Bleuler, S., Hennig, L., Prelic, A., von Rohr, P., Thiele, L., Zitzler, E., Gruissem, W., and Bühlmann, P. (2004). Sparse graphical gaussian modeling of the isoprenoid gene network in arabidopsis thaliana. *Genome Biol*, 5(11):R92.

Williams, W. (1959). Heterosis and the genetics of complex characters. *Nature*, 184:527–530.

Wissel, C. (1992). Aims and limits of ecological modelling exemplified by island theory. *Ecological Modelling*, 63:1–12.

Wolkenhauer, O. (2001). Systems biology: the reincarnation of systems theory applied in biology? *Brief Bioinform*, 2(3):258–270.

Wolkenhauer, O. (2007). Defining systems biology: an engineering perspective. *IET Syst Biol*, 1(4):204–206.

Xiao, J., Li, J., Yuan, L., and Tanksley, S. D. (1995). Dominance is the major genetic basis of heterosis in rice as revealed by QTL analysis using molecular markers. *Genetics*, 140(2):745–754.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J., and Speed, T. P. (2002). Normalization for cdna microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*, 30(4):e15.

Appendix

A Significance based network hypothesis applied to metabolite data

The experimental metabolite data described in chapter 2, were analyzed additionally in a slightly different approach (published in the proceedings Andorf et al. (2010b)). This time the network hypothesis for heterosis was based on the significance values of the calculated partial correlations, as described in chapter 3 for the gene expression profiles, instead of the partial correlation values itself.
Another difference to the study in chapter 2, is that in this analysis a linear mixed model approach was used to estimate the metabolite values for the four genotypes and seven time points. The fixed effects of this model include genotype $g \in \{\text{C24} \times \text{C24}, \text{Col-0} \times \text{Col-0}, \text{C24} \times \text{Col-0}, \text{Col-0} \times \text{C24}\}$, time point $t \in \{1, ..., 7\}$ and the interaction between genotype and time point $(g \times t)$. The model contains the random effects (underlined in the model) measuring day $\underline{d} \in \{1, ..., 3\}$ and the error random term $\underline{\varepsilon}_{i,j,k,l}$. The measuring day was used as a random effect because it is not repeatable. The fitting of the linear regression was done on a per metabolite basis for the following model where $\underline{y}_{i,j,k,l}$ depicts the logarithm of the raw metabolite signal:

$$\underline{y}_{i,j,k,l} = \mu + g_i + t_j + (g \times t)_{i,j} + \underline{d}_k + \underline{\varepsilon}_{i,j,k,l} \tag{A.1}$$

μ gives the overall metabolite-wise mean. The four genotypes are denoted with index i, the seven time points with index j, the measuring day with index k and the replicates are depicted by index l. Each genotype-time point combination was covered by four replicates. Estimated metabolite values were obtained from the linear mixed model as in Eq. A.2:

$$y^*_{i,j} = g_i + t_j + (g \times t)_{i,j} \tag{A.2}$$

Afterwards, a modest filtering step was applied on the significance of the estimated effects of the linear model where metabolites that do not show a significant (cut-

off: 0.21) time and/or genotype-time point-interaction effect are excluded from the further analysis. For this filtering step, the P-values of these effects were corrected for multiple testing using the FDR approach described by Benjamini and Hochberg (1995). After this significance filtering, 172 metabolites remained for the further analysis.

The calculation of the partial correlation values (Eq. 3.5) and partial correlation MPH (Eq. 3.9) as well as partial correlation BPH values (Eq. 3.11), based on the significance of the calculated partial correlations, was done as described for the gene expression data in chapter 3.

The histograms of the partial correlation MPH and BPH effect values for the 172 metabolites are shown in Figure A.1. The vertical lines highlight zero, where it is estimated that the metabolite is probably part of the same amount of regulatory interaction in the heterozygous genotype as, respectively, in the mean of the parents (MPH) or the better of the homozygous lines (BPH). A positive partial correlation MPH or BPH value represents that the particular metabolite is probably involved in more regulatory interactions in the hybrid than in the mid- or best-parent expectation. So, according to the network hypothesis for heterosis, it is expected that the majority of the metabolites show a positive partial correlation MPH or BPH value. This shift to positive values is visible for MPH as well as BPH of both hybrids (Figure A.1).

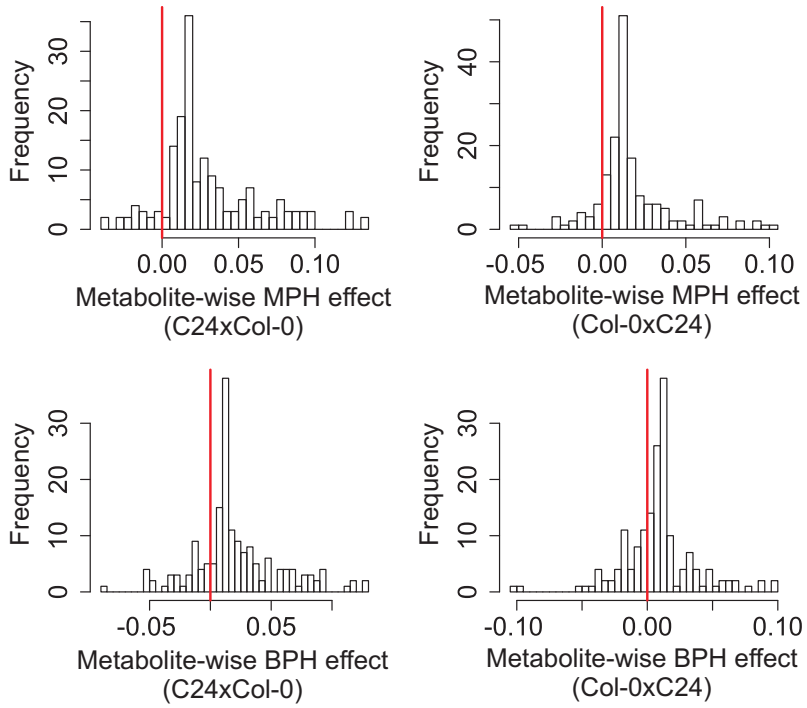


Figure A.1: Partial correlation MPH and BPH values for all metabolites after filtering with a significance cutoff of 0.21. In line with the network hypothesis for heterosis, most metabolites show positive partial correlation heterosis values.

B Influence of cutoff values in the significance filtering step

B.1 Gene expression data

In chapter 3, one particular "significance" cutoff was applied to the FDR corrected P-values of the time and genotype-time point-interaction effects from the linear model 3.3. To study the influence of the cutoff value, that is used in this filtering step, on the results of this analysis regarding the network hypothesis for heterosis, different significance cutoffs were applied to the FDR corrected P-values of the linear model 3.3. As described in chapter 3, only genes with FDR corrected P-values of the time and/or genotype-time point-interaction effects smaller than the significance cutoff were used for the further heterosis analysis. So, for small cutoffs, only genes that show a strong time dependency and/or genotype-time point-interaction were used to calculate the heterosis values according to the network hypothesis for heterosis. For large cutoff values, nearly no genes were excluded from the further analysis.
To analyze the influence of the applied significance cutoff, the following steps to calculate the heterosis values similar to the analysis in chapter 3 were done for a set of different significance cutoffs (from 0.001 to 1 by steps of 0.01):

1. Applying the current significance cutoff value to the FDR corrected P-values.

2. Choosing 1000 genes randomly out of the set of "significant" genes (genes remaining after the particular significance cutoff was applied) for five times.

3. Calculating the partial correlation heterosis values for each of the five sets of 1000 genes according to Eqs. 3.9 for MPH and 3.11 for BPH.

4. Computing for any of the five sets of significant genes the median values of all these heterosis values, separately for the two hybrids and MPH as well as BPH.

5. Determining separately for either hybrid and both heterosis measures, the mean of the median values. So, one mean value is calculated for each hybrid-heterosis measure combination.

In Figure B.1, these mean values are plotted against the significance cutoff values. For all cutoffs larger than a certain value between 0.16 and 0.26, depending on the hybrid and heterosis measure, the average of the median of the heterosis values according to the network hypothesis for heterosis is positive. A positive mean of the

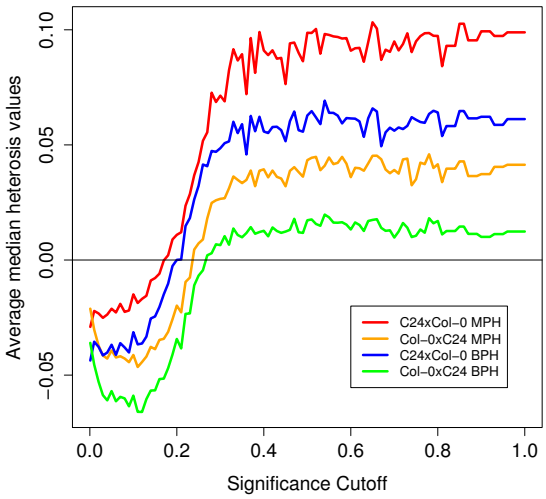


Figure B.1: The filtering step of applying a significance cutoff to FDR corrected P-values of time and genotype-time point-interaction effects of the linear model 3.3 has an influence on the decision if the network hypothesis for heterosis has to be rejected or not. For small significant cutoffs, the hypothesis does not hold true (negative average median heterosis values).

five median values is observed if the majority of genes of the particular set of genes show a positive heterosis effect regarding the network hypothesis for heterosis. In other words, in case of a positive mean, it is estimated that the majority of genes is involved in more regulatory interactions in the hybrid than in the mid-parent or best-parent expectation.

B.2 Metabolite data

For the metabolite profiles (chapter 2), the same study, as described in the previous section for the gene expression data, of the influence of the significance cutoff was performed (Andorf et al., 2010b). The basis for this study built the FDR corrected P-values of the effects in the linear mixed model A.1 described in appendix A. As in the analysis of the gene expression data, different significance cutoffs (from 0.001 to 1 by steps of 0.001) were applied to exclude metabolites from the respective analysis that show larger corrected P-values of their time and genotype-time point-interaction effect than the used significance cutoff. Different from the study of the gene expression profiles, where five samples of 1000 genes each were selected and analyzed, the partial correlation heterosis values could be calculated at once for all 192 metabolites under study for each significance cutoff. So, for every significant

cutoff in this screening analysis, the partial correlation MPH and BPH values for either hybrid were calculated and in each case the median of the partial correlation heterosis values was determined. These median values were plotted against the significance cutoffs (Figure B.2). The outcome of this study is the same as for the gene expression data. Too small cutoff values lead to no positive heterosis values that were expected in our hypothesis. For large significance cutoffs, the network hypothesis for heterosis holds true for either hybrid regarding both heterosis measures.

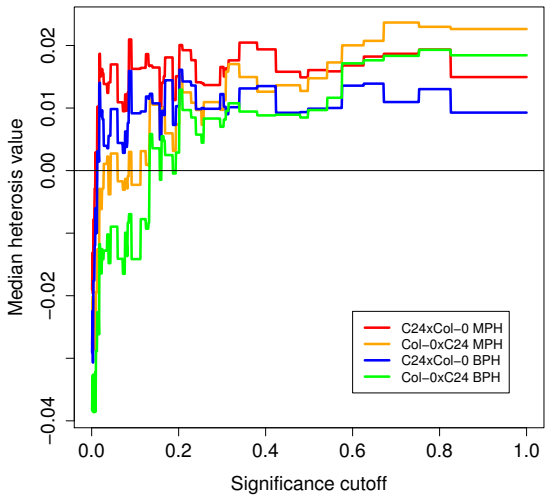


Figure B.2: The median of the partial correlation MPH and BPH values are plotted against different cutoffs for the significance filter for the time and/or genotype-time point-interaction effects after the metabolite-wise linear mixed model A.1. Only for small significance cutoff values, the network hypothesis for heterosis has to be rejected (negative median heterosis values).

Allgemeinverständliche Zusammenfassung

Als Heterosis-Effekt wird die Überlegenheit in einem oder mehreren Leistungsmerkmalen (z.B. Blattgröße von Pflanzen) von heterozygoten (mischerbigen) Nachkommen über deren unterschiedlich homozygoten (reinerbigen) Eltern bezeichnet. Dieses Phänomen ist schon seit Beginn des letzten Jahrhunderts bekannt und wird weit verbreitet in der Pflanzenzucht genutzt. Trotzdem sind die genetischen und molekularen Grundlagen von Heterosis noch weitestgehend unbekannt.

Es wird angenommen, dass heterozygote Individuen mehr regulatorische Möglichkeiten aufweisen als ihre homozygoten Eltern und sie somit auf eine größere Anzahl an wechselnden Umweltbedingungen richtig reagieren können. Diese erhöhte Anpassungsfähigkeit führt zum Heterosis-Effekt.

In dieser Arbeit wird ein systembiologischer Ansatz, basierend auf molekularen Netzwerkstrukturen verfolgt, um zu einem besseren Verständnis von Heterosis beizutragen. Dazu wird eine Netzwerkhypothese für Heterosis vorgestellt, die vorhersagt, dass die heterozygoten Individuen, die Heterosis zeigen, mehr regulatorische Interaktionen in ihren molekularen Netzwerken aufweisen als die homozygoten Eltern. Partielle Korrelationen wurden verwendet, um diesen Unterschied in den globalen Interaktionsstrukturen zwischen den Heterozygoten und ihren homozygoten Eltern zu untersuchen.

Die Netzwerkhypothese wurde anhand von Metabolit- und Genexpressionsdaten der beiden homozygoten *Arabidopsis thaliana* Pflanzenlinien C24 und Col-0 und deren wechselseitigen Kreuzungen getestet. *Arabidopsis thaliana* Pflanzen sind bekannt dafür, dass sie einen Heterosis-Effekt im Bezug auf ihre Biomasse zeigen. Die heterozygoten Pflanzen weisen bei gleichem Alter eine höhere Biomasse auf als die homozygoten Pflanzen.

Die Netzwerkhypothese für Heterosis konnte sowohl im Bezug auf mid-parent Heterosis (Unterschied in der Leistung des Heterozygoten im Vergleich zum Mittelwert der Eltern) als auch auf best-parent Heterosis (Unterschied in der Leistung des Heterozygoten im Vergleich zum Besseren der Eltern) für beide Kreuzungen für die Metabolit- und Genexpressionsdaten bestätigt werden.

In einer Überrepräsentations-Analyse wurden die Gene, für die die größte Veränderung in der Anzahl der regulatorischen Interaktionen, an denen sie vermutlich beteiligt sind, festgestellt wurde, mit den Genen aus einer quantitativ genetischen (QTL) Analyse von Biomasse-Heterosis in *Arabidopsis thaliana* verglichen. Die ermittelten Gene aus beiden Studien zeigen eine größere Überschneidung als durch Zufall erwartet. Das deutet darauf hin, dass jede identifizierte QTL-Region viele Gene, die den Biomasse-Heterosis-Effekt in *Arabidopsis thaliana* beeinflussen, enthält. Die Gene, die in den Ergebnislisten beider Analyseverfahren überlappen, können mit größerer Zuversicht als Kandidatengene für Biomasse-Heterosis in *Arabidopsis thaliana* betrachtet werden als die Ergebnisse von nur einer Studie.

Acknowledgment

First of all I would like to thank my supervisor, Dirk Repsilber, for supporting my work in all possible ways. He provided me with a lot of important advice and helpful suggestions. Additionally, I wish to express my sincere appreciation to him for his untiring encouragement during the course of this work.

Thanks to Joachim Selbig for mentoring this work and his continual help.

Many thanks go to Rhonda Meyer and Thomas Altmann for their continual help and the great and fruitful collaboration. Without their QTL data and help in working with it, one important part of this work would have been impossible. Special thanks to Rhonda for helping a lot with my understanding of QTL and for never getting tired of answering my questions.

I am very grateful to Nina Melzer, Dörte Wittenburg, Daisy Zimmer, Vinzent Börner and Nadine Neugebauer for all their help, support and patience during the whole project. Furthermore, I would like to thank all present and former members of the Genetics and Biometry department for their advice and support in performing this study.

I sincerely thank the whole Biostatistics Group at the Norwegian University of Life Sciences for some great and inspiring months in Norway.

Additional thanks go to Tanja Gärtner and Matthias Steinfath for many helpful tips and suggestions.

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig und unter Verwendung keiner anderen als den von mir angegebenen Quellen und Hilfsmitteln verfasst habe.
Ferner erkläre ich, dass ich bisher weder an der Universität Potsdam noch anderweitig versucht habe, eine Dissertation einzureichen oder mich einer Doktorprüfung zu unterziehen.

Sandra Andorf
Potsdam, den 09.11.2010

Publications

1. S. Andorf, T. Gärtner, M. Steinfath, H. Witucka-Wall, T. Altmann, and D. Repsilber. (2009) *Towards Systems Biology of Heterosis: A Hypothesis about Molecular Network Structure Applied for the Arabidopsis Metabolome.* EURASIP J Bioinform Syst Biol, articleID: 147157

2. S. Andorf, J. Selbig, T. Altmann, K. Poos, H. Witucka-Wall, and D. Repsilber. (2010) *Enriched partial correlations in genome-wide gene expression profiles of hybrids (A. thaliana): a systems biological approach towards the molecular basis of heterosis.* Theor Appl Genet, 120(2):249-259

3. S. Andorf, R.C. Meyer, J. Selbig, T. Altmann, and D. Repsilber. (2010) *Integration of a systems biological network analysis and QTL results for biomass heterosis in Arabidopsis thaliana.* Under review

1-3: SA contributed to concept and method development, data analysis and preparation of the manuscript

4. T. Gärtner, M. Steinfath, **S. Andorf**, J. Lisec, R.C. Meyer, T. Altmann, L. Willmitzer, and J. Selbig. (2009) *Improved Heterosis Prediction by Combining Information on DNA- and Metabolic Markers.* PLoS One, 4(4): e5220

4: SA contributed to a database analysis based on KEGG

Résumé

This page contains personal information and is, therefore, excluded from the online publication.